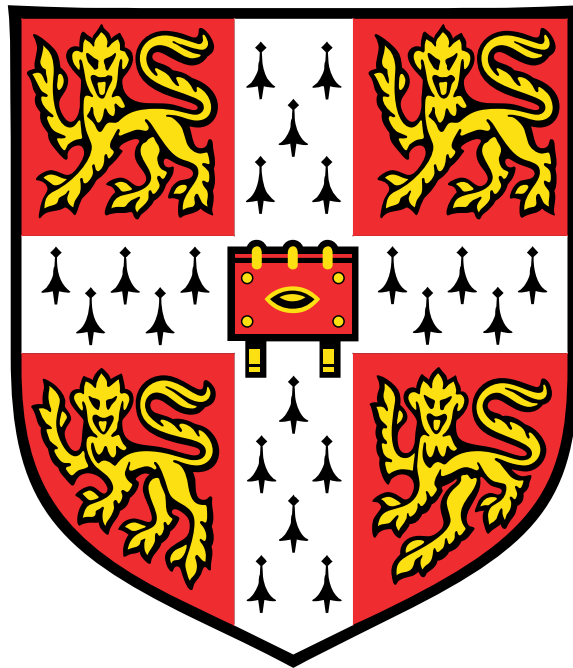


The pathological and genomic impact of CTCF depletion in mammalian model systems



Sarah Jane Aitken

CRUK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College

August 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other, university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding appendices, references, tables, and equations and has fewer than 150 figures.

Sarah Jane Aitken
August 2018

Acknowledgements

I would like acknowledge the people who have contributed to this work and to thank my many new friends and colleagues.

First, thank you to my supervisor, Duncan Odom, for the opportunity to carry out my PhD in his laboratory, providing close mentorship throughout my training, and offering valuable scientific freedom. I also thank my fellow Odom lab members for the stimulating scientific working environment and being supportive colleagues.

This work would not have been possible without the time and efforts of several colleagues and collaborators: Ximena Ibarra Soria, with whom I worked closely and who performed a significant part of the integrative analyses presented in Chapter 3, and John Marioni for her supervision; Christine Feig, for scientific guidance and validation of the genetically engineered mouse lines; Tim Rayner, for the whole genome sequencing analyses presented in Chapter 4; the other Liver Cancer Evolution team members, Frances Connor and Margus Lukk; Oriol Pich (IRB Barcelona) for the RNA-sequencing and motif analyses presented in Chapter 4, and Núria López-Bigas for his supervision; Elsa Kentepozidou (EMBL-EBI) for providing the unpublished data included in Chapter 3; Paul Flicek (EMBL-EBI) for insightful discussions; and finally, Susan Davies (Addenbrooke's Hospital) for liver pathology advice.

The technical breadth of this project would not have been possible without the excellent core facilities at the CRUK Cambridge Institute, especially Angela Mowbray in the Biological Resources Unit. The Institute has also provided an outstanding scientific environment and has given me the opportunity to collaborate on other projects, including with Christina Ernst, Isabel Quirós-González, and Matthew Hoare, providing me with different and valuable perspectives.

I am extremely grateful for the personal funding provided by the Wellcome Trust Programme for Clinicians Fellowship; an EMBO|EuropaBio Genomics & Computational Biology Visiting Fellowship; the Pathological Society for two Trainee Research Grants; and Cancer Research UK for core funding of the Cambridge Institute.

Finally, to my partner, Dana, for his endless support, understanding, and encouragement in my academic and personal endeavours.

Abstract

CCCTC-binding factor (CTCF) binds DNA, thereby helping to partition the mammalian genome into discrete structural and regulatory domains. In doing so, it insulates chromatin and fine-tunes gene activation, repression, and silencing. Complete removal of CTCF from mammalian cells causes catastrophic genomic dysregulation, most likely due to widespread collapse of 3D chromatin looping within the nucleus. In contrast, *Ctcf* hemizygous mice with lifelong reduction in CTCF expression are viable but have an increased incidence of spontaneous multi-lineage malignancies. In addition, *CTCF* is mutated in many human cancers and is thus implicated as a tumour suppressor gene. This study aimed to interrogate the genome-wide consequences of a reduced genomic concentration of *Ctcf* and its implications for carcinogenesis.

In a genetically engineered mouse model, *Ctcf* hemizygous cells showed modest but robust changes in almost a thousand sites of genomic CTCF occupancy; these were enriched for lower affinity binding events with weaker evolutionary conservation across the mouse lineage. Furthermore, several hundred genes concentrated in cancer-related pathways were dysregulated due to changes in transcriptional regulation. Global chromatin structure was preserved but some loop interactions were destabilised, often around differentially expressed genes and their enhancers. Importantly, these transcriptional alterations were also seen in human cancers.

These findings were then examined in a hepatocyte-specific mouse model of *Ctcf* hemizygosity with diethylnitrosamine-induced liver tumours. *Ctcf* hemizygous mice had a subtle liver-specific phenotype, although the overall tumour burden in *Ctcf* hemizygous and wild-type mice was the same. Using whole genome sequencing, the highly reproducible mutational signature caused by DEN exposure was characterised, revealing that *Braf*(V637E), orthologous to *BRAF*(V600E) in humans, was the predominant oncogenic driver in these liver tumours.

Taken together, while *Ctcf* loss is partially physiologically compensated, chronic CTCF depletion dysregulates gene expression by subtly altering transcriptional regulation. This study also represents the first comprehensive genome-wide and histopathological characterisation of this commonly used liver cancer model.

Table of contents

List of figures	xiii
List of tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Mammalian gene regulation	1
1.1.1 Epigenetics	3
1.2 CTCF is a multi-functional DNA-binding protein	7
1.2.1 CTCF in development, imprinting, and X inactivation	9
1.2.2 Transcriptional regulation	10
1.2.3 CTCF-dependent chromatin looping	11
1.3 Cancer epigenetics	13
1.3.1 <i>CTCF</i> is a tumour suppressor gene	16
1.3.2 <i>Ctcf</i> haploinsufficiency predisposes to cancer in mice	18
1.3.3 Mutation of <i>CTCF</i> and its binding sites in human cancers	18
1.4 Mouse models of liver cancer	20
1.5 Thesis outline	23
2 Materials and methods	25
2.1 Mouse colony management	25
2.1.1 Genetically engineered mice	25
2.1.2 Mouse genotyping	26
2.2 Mouse embryonic fibroblast cultures	26
2.3 qPCR	27
2.4 Quantitative western blotting	28
2.5 ChIP-sequencing	28
2.5.1 Computational analyses of ChIP-sequencing	32

2.6	RNA-sequencing	35
2.6.1	Computational analyses of RNA-sequencing	38
2.7	TMT proteomics	40
2.7.1	Computational analyses of proteomic data	42
2.8	Hi-C	42
2.8.1	Computational analyses of Hi-C data	46
2.9	Mouse tumour models	48
2.9.1	Spontaneous tumourigenesis	48
2.9.2	Chemical model of hepatocarcinogenesis	48
2.10	Tissue collection and processing	49
2.10.1	Fresh frozen tissue	49
2.10.2	Fixed tissue for histology	50
2.11	Tumour histopathology	51
2.12	Whole genome sequencing	52
2.12.1	Computational analyses of whole genome sequencing	53
2.13	Data storage and management	56
3	CTCF maintains regulatory homeostasis of cancer pathways	59
3.1	Introduction	59
3.1.1	Project aim and overview	62
3.2	Results	63
3.2.1	Successful generation of embryonic fibroblast cultures	63
3.2.2	Molecular characterisation of <i>Ctcf</i> hemizygous MEFs	65
3.2.3	Chronic reduction of CTCF alters its chromatin binding	66
3.2.4	Labile CTCF binding sites have distinct genomic features	67
3.2.5	<i>Ctcf</i> hemizygosity alters transcription of cancer pathways	67
3.2.6	Gene expression changes correspond with altered looping	70
3.2.7	Altered gene expression patterns are found in human tumours	74
3.3	Discussion	76
4	Genetic and chemical models of hepatocarcinogenesis	79
4.1	Introduction	79
4.1.1	Project aim	81
4.2	Results	82
4.2.1	Characterisation of <i>Ctcf</i> hemizygous mice	82
4.2.2	Hepatocyte-specific <i>Ctcf</i> knockdown	90
4.2.3	Tumour induction using diethylnitrosamine	94

4.2.4	Pathological characterisation of DEN-induced tumours	96
4.2.5	Genomic characterisation <i>Ctcf</i> hemizygous liver tumours . .	100
4.2.6	Distinct mutational signatures of liver tumours	102
4.2.7	<i>Braf</i> is the predominant driver of DEN-induced liver tumours	106
4.2.8	<i>Apc</i> is a secondary driver in mouse HCC	110
4.2.9	CTCF binding sites are enriched for mutations	110
4.2.10	<i>Carmil2</i> is overexpressed in <i>Ctcf</i> hemizygous tumours	113
4.3	Discussion	115
5	Discussion and outlook	121
5.1	CTCF haploinsufficiency	121
5.2	Model systems	122
5.3	Mechanisms of chromatin organisation	125
5.4	DNA damage and repair	127
	Publications	131
	References	133
	Appendix A: List of differentially expressed genes	165
	Appendix B: List of significantly mutated genes	171

List of figures

1.1	Eukaryotic transcriptional regulation	2
1.2	Regulation of chromatin by histone modifications	5
1.3	Higher-order chromatin structure	8
1.4	Models of TAD formation	13
3.1	Project overview	62
3.2	Conditional deletion of the mouse <i>Ctcf</i> gene	63
3.3	MEF cultures	64
3.4	Validation of CTCF depletion	65
3.5	<i>Ctcf</i> hemizyosity results in altered chromatin binding	66
3.6	Differentially bound CTCF loci have distinct genomic features	68
3.7	CTCF depletion dysregulates oncogenic pathways	69
3.8	Transcriptional perturbations arise from regulatory changes in the nuclear genome	71
3.9	Global-scale chromatin interactions are robust to reduced CTCF levels	73
3.10	Concordant gene alterations in diverse murine and human tumours .	75
4.1	Histology of <i>Ctcf</i> hemizygous mice	84
4.2	Histological characterisation of hepatocellular neoplasms	89
4.3	Hepatocyte-specific <i>Ctcf</i> knockdown.	93
4.4	Overview of tumour induction protocol	94
4.5	DEN-initiated tumour characteristics	96
4.6	Histology of hepatocellular neoplasms	97
4.7	Macroscopic and microscopic appearance of liver tumours	99
4.8	Independent evolution of DEN-induced tumours	101
4.9	DEN-initiated neoplasms have a high SNV burden	103
4.10	DEN-initiated tumours have distinct mutational signatures	105
4.11	Mutational signatures of DEN-induced and spontaneous tumours . .	106

4.12 DEN-initiated tumours carry <i>Braf</i> and <i>Hras</i> mutations	108
4.13 Validation of <i>Apc</i> mutations	111
4.14 CTCF binding affinity correlates with mutational burden	112
4.15 CTCF binding sites are hypermutated	113
4.16 Differential gene expression in liver and DEN-induced tumours . . .	114

List of tables

2.1	Primers used for mouse genotyping	27
2.2	Library amplification cycles	31
2.3	Antibodies used for immunohistochemistry	51
3.1	Genotyping of mouse embryos	64
4.1	Genotyping of <i>Ctcf</i> hemizygous mice	82
4.2	Prevalence of spontaneous tumours in aged <i>Ctcf</i> hemizygous mice	87
4.3	Genotyping of liver-specific <i>Ctcf</i> hemizygous mice	91
4.4	Pilot experiment to determine time points for sample collection . . .	95
4.5	Hotspot mutations in oncogenes and tumour suppressor genes . . .	109

Abbreviations

AATK	apoptosis-associated tyrosine kinase
ALD	alcoholic liver disease
APC	adenomatous polyposis coli
AR	androgen receptor
ATP	adenosine triphosphate
AWERB	Animal Welfare and Ethical Review Body
BCL2	B-cell lymphoma 2
BCL6	B-cell lymphoma 6
BER	base excision repair
bp	base pair
BRAF	B-Raf proto-oncogene, serine/threonine kinase
bRP	basic reverse phase
BSA	bovine serum albumin
Carmil2	capping protein, Arp2/3 and myosin-I linker protein 2
Cas	CRISPR associated protein
CC	cholangiocarcinoma
ChIP-seq	chromatin immunoprecipitation followed by high-throughput sequencing
cis-SAGe	cis-splicing between adjacent genes
CNV	copy number variation
CRC	colorectal cancer
CRISPR	clustered regularly interspaced short palindromic repeats

CTCF	CCCTC-binding factor
DAB	3,3'-diaminobenzidine
DEN	diethylnitrosamine
DMD	differentially methylated domain
DMEM	Dulbecco's Modified Eagle medium
DN	dysplastic nodule
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
DPX	distyrene plasticiser xylene
E	embryonic day
EDTA	ethylenediaminetetraacetic acid
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EGTA	ethylene glycol-bis(2-aminoethylether)-tetraacetic acid
EMH	extramedullary haematopoiesis
emRiboSeq	embedded ribonucleotide sequencing
EMT	epithelial-mesenchymal transition
EndoSeq	endonuclease sequencing
ERG	ETS-related gene
ETS	erythroblast transformation-specific
EWAS	epigenome-wide association study
F1	filial generation one
FACS	fluorescence activated cell sorting
FBS	foetal bovine serum
Fbxo6	F-box only protein 6
FDR	false discovery rate

FF	fresh frozen
FFPE	formalin-fixed paraffin-embedded
FOXA1	forkhead box A1
G&T-seq	genome and transcriptome sequencing
gDNA	genomic DNA
GEM	genetically-engineered mouse
GRCh38	Genome Reference Consortium human build 38
GRCm38	Genome Reference Consortium mouse build 38
GTF	general transcription factor
GWAS	genome-wide association study
H3K27ac	acetylation of histone H3 at lysine 27
H3K27me3	trimethylation of histone H3 at lysine 27
H3K4me1	monomethylation of histone H3 at lysine 4
H3K4me3	trimethylation of histone H3 at lysine 4
H3K9me3	trimethylation of histone H3 at lysine 9
H&E	haematoxylin and eosin
HAT	histone acetyltransferase
HBV	hepatitis B virus
HCC	hepatocellular carcinoma
HCV	hepatitis C virus
HDAC	histone deacetylase
HE	hemizygous
Hi-C	high-resolution chromosome conformation capture sequencing
HIER	heat induced epitope retrieval
hMLH1	human MutL homologue 1
HNF4A	hepatocyte nuclear factor 4 alpha

HPLC	high-performance liquid chromatography
HRAS	Harvey rat sarcoma viral oncogene homologue
Hz	Hertz
ICGC	International Cancer Genome Consortium
IDH1	isocitrate dehydrogenase 1
IGF2	insulin-like growth factor 2
IHC	immunohistochemistry
IL-6	interleukin 6
INHAND	International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice
IP	intraperitoneal
IQR	interquartile range
KRAS	Kirsten rat sarcoma viral oncogene homologue
LB	lysis buffer
LC-MS	liquid chromatography-mass spectrometry
LCR	locus control region
LICA-FR	liver cancer (France)
lncRNA	long noncoding RNA
LOF	loss of function
LOH	loss of heterozygosity
MEF	mouse embryonic fibroblast
MGMT	O ⁶ -methylguanine-DNA methyltransferase
miRNA	micro RNA
MMS	methyl methanesulfonate
MMTS	methyl methanethiosulfonate
Mnd1	meiotic nuclear division protein 1
MRN	macroregenerative nodule

MS	mass spectrometry
mTOR	mammalian target of rapamycin
NAFLD	non-alcoholic fatty liver disease
ncRNA	noncoding RNA
NER	nucleotide excision repair
NGS	next-generation sequencing
NIPBL	nipped-B-like protein
NP40	Nonidet P-40
NRAS	neuroblastoma RAS viral oncogene homologue
nt	nucleotide
NTB	no Tween buffer
Nudt11	diphosphoinositol polyphosphate phosphohydrolase 3-beta
ORO	oil red O
PBS	phosphate-buffered saline
PCA	principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCI	phenol:chloroform:isoamyl alcohol
PCR	polymerase chain reaction
PGK	phosphoglycerate kinase 1
PGR	progesterone
PI	protease inhibitor
PIER	proteolytic induced epitope retrieval
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
piRNA	PIWI-interacting RNA
PLG	phase lock gel
Pol II	polymerase II

PTEN	phosphatase and tensin homologue
QC	quality control
RB1	retinoblastoma 1
rcf	relative centrifugal force
RIPA	radioimmunoprecipitation assay
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RNP	ribonucleoprotein
ROS	reactive oxygen species
RSB	resuspension buffer
RT	room temperature
scNMT-seq	single-cell nucleosome, methylation, and transcription sequencing
scRNA-seq	single-cell RNA sequencing
SDS	sodium dodecyl sulfate
siRNA	small interfering RNA
SMC	structural maintenance of chromosomes
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SPRI	solid phase reversible immobilisation
SPS	synchronous precursor selection
SV40	simian vacuolating virus 40
TAD	topologically associated domain
TATA	TATAAATA
TB	Tween buffer
TBS	Tris-buffered saline

TCEP	Tris(2-carboxyethyl)phosphine
TCGA	The Cancer Genome Atlas
TE	Tris-EDTA
TEAB	triethylammonium bicarbonate
TERT	telomerase reverse transcriptase
tet	tetracycline
TFIID	transcription factor II D
TGF- α	transforming growth factor alpha
TIL	tumour infiltrating lymphocyte
TLE	Tris-low-EDTA
TMT	tandem mass tag
TP53	tumour protein p53
TPM	transcripts per million reads
TSG	tumour suppressor gene
TSS	transcription start site
VAF	variant allele frequency
WAPL	wings apart-like protein
WES	whole exome sequencing
WGS	whole genome sequencing
WT	wild-type
XIST	X-inactive specific transcript

Chapter 1

Introduction

1.1 Mammalian gene regulation

The phenotype-genotype gap – the inability to explain phenotypic diversity on the basis of DNA sequence alone – is apparent at inter-cellular, inter-individual, and inter-species levels. Every cell within an individual's tissues and organs shares the same genetic sequence, yet each is functionally and morphologically distinct; for example, hepatocytes are relatively large cells measuring 20-30 μm in diameter specialised for synthesis, metabolism, and detoxification, while lymphocytes are 7 μm in diameter and mediate immune processes. At the population level, <0.1% of bases are different between individual humans (The 1000 Genomes Project Consortium, 2012), and although 96% of our genes are homologous with chimpanzees (*Pan troglodytes*) (The Chimpanzee Sequencing and Analysis Consortium, 2005; Varki and Altheide, 2005) and over 35% are homologous with worms (such as *Caenorhabditis elegans*) (C. elegans Sequencing Consortium., 1998) or the common fruit fly (*Drosophila melanogaster*) (Adams et al., 2000), phenotypic differences are obvious. Phenotype is not simply a product of DNA sequence alone, and many complex extra-genomic and epigenetic processes contribute to phenotypic diversity. The most important of these is the regulation of gene expression.

The structural and physiological complexities of multicellular eukaryotes demand more intricate gene regulation than that described in prokaryotes (Jacob and Monod, 1961). The first novel mechanism of transcriptional regulation to emerge in eukaryotes was that of *trans*-acting sequence-specific transcription factors binding to *cis*-regulatory DNA sequences to regulate transcription by RNA polymerases (Roeder and Rutter, 1969). The default eukaryotic transcriptional state is “off”, with

gene expression requiring combinatorial binding of multiple transcription factors to gene promoter or enhancer regions to coordinate transcription (Reményi et al., 2004). This fundamental control process is fine-tuned by a number of mechanisms that are not strictly dependent on protein-coding DNA sequences – termed epigenetic mechanisms – including DNA methylation, histone modification, expression of noncoding (nc)RNAs, and higher-order chromatin structure (**Figure 1.1**). These processes that influence DNA function and their molecular and cellular effects have given rise to the field of epigenetics.

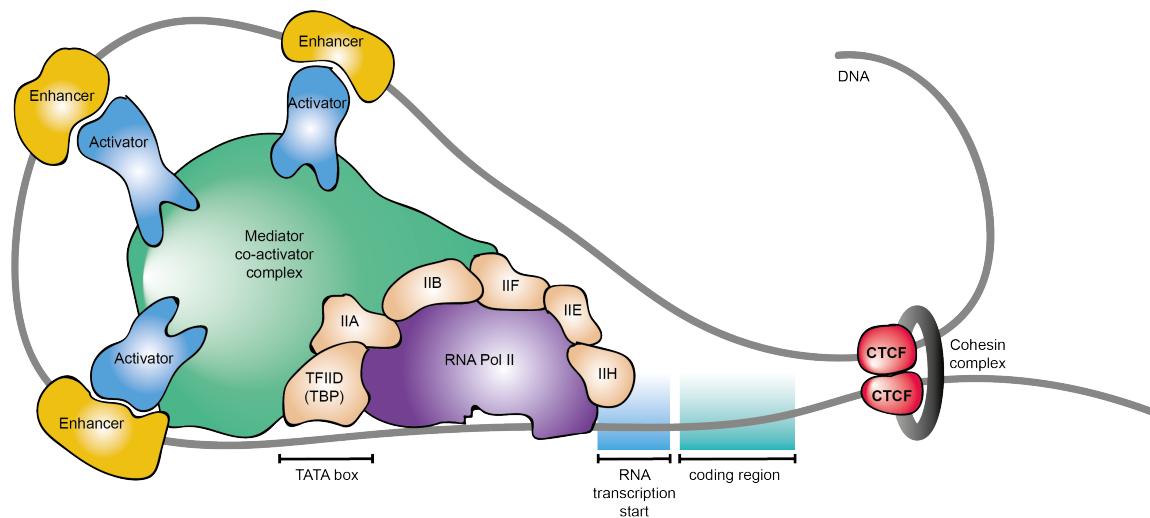


Fig. 1.1 Eukaryotic transcriptional regulation. DNA looping mediated by CCCTC binding factor (CTCF) brings distal enhancer elements into close proximity to promoter elements. The promoter is the DNA sequence where the general transcription factors (GTFs) and polymerase assemble. These elements are immediately upstream of protein-coding genes marked by active histone marks such as H3K4me3 and unmethylated CpG islands, and are bound by tissue-specific transcription factors that recruit GTFs and the RNA polymerase II (Pol II) machinery. Enhancers are *cis*-regulatory DNA sequences that are marked by active histone modifications such as H3K4me1 and are bound by tissue-specific transcription factors and transcriptional co-activators such as p300, thus affecting the rate of transcription. These sequences can locate immediately adjacent to the promoter, far upstream of it, within introns, or entirely downstream of the gene. The length of DNA between the *cis*-regulatory sequences and the start of transcription varies, sometimes reaching tens of thousands of nucleotides in length. Before the start of transcription, the transcription factor IID (TFIID) complex binds to the TATA box in the core promoter of the gene. Many transcriptional regulators act through Mediator (a multi-protein complex that functions as a transcriptional co-activator in all eukaryotes (Kelleher et al., 1990)), while some interact with GTFs and RNA Pol II directly. Transcriptional regulators also act by recruiting proteins that alter the chromatin structure of the promoter. Whereas Mediator and the GTFs are the same for many Pol II-transcribed genes, the transcriptional regulators and the locations of their binding sites relative to the promoter differ for each gene (Alberts et al., 2014).

1.1.1 Epigenetics

Epigenetic traits are defined as “stably heritable phenotypes resulting from changes in a chromosome without alterations in the DNA sequence” (Berger et al., 2009). This definition has markedly evolved since Waddington originally fused the words “epigenesis” (cell differentiation) and “genetics” in the context of developmental biology (Waddington, 1942). His epigenetic landscape metaphor detailed how gene regulation modulates embryonic development and how mutations alter cell fates and consequently morphogenesis. Over a decade later, Nanney (1958) adopted the term “epigenetics” to describe auxiliary mechanisms that determine which genes are expressed in a particular cell. The definition then stabilised to describe “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence” (Riggs and Porter, 1996). More recently, others have included “transgenerational inheritance” of epigenetic marks such as DNA methylation or histone modifications via gametes, first described in plants but now also well recognised in mammals (Anway et al., 2005) including humans (Skinner, 2014), within the umbrella of epigenetics.

Understanding the mechanisms involved in the initiation, maintenance, and heritability of epigenetic states is an important aspect of current biological research, since these mechanisms underpin how each cell behaves in the complex system of the organism. Epigenetics is now known to contribute to several major pathologies including cancer, chromosomal instabilities, and mental retardation (Egger et al., 2004). To understand the ways in which epigenetic control contributes to cancer initiation and development, it is first necessary to be familiar with the main epigenetic mechanisms.

Methylation

DNA methylation (addition of a -CH₃ group) at the C5 position of DNA cytosine residues is the archetypal epigenetic gene control mechanism (Holliday and Pugh, 1975) and typically has a gene silencing effect (Egger et al., 2004). Methylation is catalysed by DNA methyltransferases (DNMTs) and, in mammals, typically occurs at symmetrical CpG dinucleotides. Methylation patterns are defined during embryological development and retained during somatic cell division (Messerschmidt et al., 2014). CpG islands, which are regions ≥ 500 bp in length with a GC content $>55\%$ present at the 5' promoter region of approximately 50% of genes (Takai and Jones, 2002), are spared methylation and tend to be transcriptionally active housekeeping

genes. *De novo* methylation of these CpG islands may result in aberrant gene expression such as the silencing of tumour suppressor genes in cancer cells (Jones and Baylin, 2002) .

Histone modifications

DNA wraps around histone proteins to form nucleosomes, the basic unit of chromatin and the means by which DNA is packaged in eukaryotic cells. The acetylation, phosphorylation, or methylation of histone octamers (tetramer of two histone 2A (H2A) and two histone 2B (H2B) molecules and H3 and H4 dimers) allow dynamic regulation of chromatin and, therefore, gene expression (Taby and Issa, 2010) (**Figure 1.2A**).

Acetylation of lysines is highly dynamic and regulated by the opposing action of two families of enzymes, histone acetyltransferases (HATs) and histone deacetylases (HDACs). HATs use acetyl CoA as a cofactor and catalyse the transfer of an acetyl group to the ϵ -amino group of lysine side chains, thereby neutralising lysine's positive charge; this action has the potential to weaken histone-DNA interactions (Parthun, 2007). De-acetylation of the positively charged N-termini of H3 and H4 results in a closed and tight chromatin configuration (heterochromatin) around the negatively charged DNA, whereas acetylation neutralises this charge and chromatin becomes more open (euchromatin) and, therefore, more permissible to transcription (Struhl, 1998) (**Figure 1.2B**). Histone phosphorylation is also highly dynamic and takes place on serines, threonines and tyrosines, predominantly, but not exclusively, at N-terminal histone tails. Phosphorylation levels are controlled by kinases and phosphatases that add or remove the modification, respectively (Oki et al., 2007).

Methylation mainly occurs on the side chains of lysines and arginines. Unlike acetylation and phosphorylation, histone methylation does not alter the histone protein charge. The epigenetic effect of histone methylation is more variable than that of acetylation, in part because lysines may be mono-, di-, or tri-methylated and arginines may be mono-, symmetrically, or asymmetrically di-methylated; the net result is that regions can be marked as either active or inactive. For example, H3K4me3 (trimethylation of lysine 4 of histone H3) occurs at active gene promoter regions (Lachner and Jenuwein, 2002), whereas DNA methylation and repressive histone marks such as H3K9me3 and H3K27me3 mark silent heterochromatin that is not being actively transcribed such as centromeres and telomeres (Zhang et al., 2015). Repeat elements are mostly methylated to prevent their activity (Bedford and Clarke, 2009).

Finally, ubiquitination results in a much larger covalent modification than acetylation, phosphorylation, or methylation. Ubiquitin is a 76-amino acid polypeptide that attaches to histone lysines via the sequential action of three enzymes E1, E2, and E3, which perform activation, conjugation, and ligation, respectively. Histones usually become mono-ubiquitylated, and the modification is removed via the action of isopeptidases called deubiquitinating enzymes, which is important for both gene activity and silencing (Wang et al., 2004).

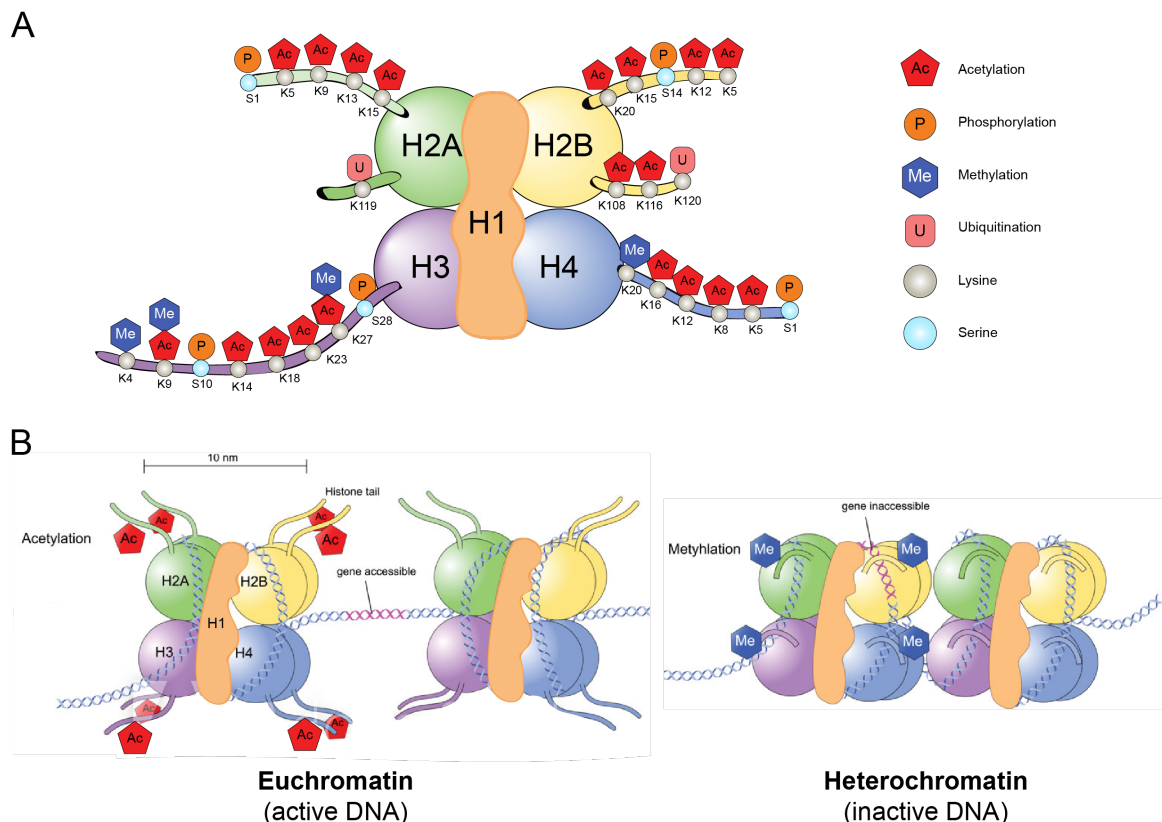


Fig. 1.2 Regulation of chromatin by histone modifications. Histone modifications play fundamental roles in most biological processes involved in the manipulation and expression of DNA including gene transcription (Bannister and Kouzarides, 2011). (A) Examples of proteins domains that specifically bind to modified histones are shown (Kouzarides, 2007). (B) Euchromatin and heterochromatin are biochemically distinguishable by different covalent histone modifications: de-acetylation of H3 and H4 results in a closed and tight chromatin configuration (heterochromatin) around negatively charged DNA, whereas acetylation neutralises this charge and chromatin becomes more open (euchromatin) and, therefore, more permissible to transcription (Struhl, 1998).

Noncoding RNAs

The vast majority of genomic regions are transcribed into mRNA, but only 2% are ultimately translated into protein (Liu et al., 2013a). These non-protein coding RNA (ncRNA) transcripts are important epigenetic gene regulatory mechanisms (Liu et al., 2013a). ncRNAs are subclassified according to size into long ncRNAs (lncRNAs, >200 nucleotides (nt)) or short ncRNAs (<200 nt), the latter including microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs) (Peschansky and Wahlestedt, 2014).

Small ncRNAs are responsible for RNA gene silencing, with each type exploiting independent mechanisms of action. For example, piRNAs preserve mammalian genome integrity by regulating the activity of transposable elements (Ernst et al., 2017). lncRNAs are far more diverse in structure and function and may behave as signals, decoys, guides, or scaffolds (Wang and Chang, 2011) to form ribonucleic-protein (RNP) complexes with other chromatin regulators that coordinate the regulation of gene expression (Rinn and Chang, 2012). They can be broadly classified into (i) those acting in *cis* to control expression and/or the chromatin state of nearby genes, and (ii) those which act in *trans* to cause a diverse array of functions throughout the cell. At least three proposed mechanisms for gene expression regulation are proposed for *cis*-acting lncRNAs: (i) the lncRNA transcript itself regulates the expression of neighbouring genes via its ability to recruit regulatory factors to the locus and/or modulate their function; (ii) the process of transcription and/or splicing of the lncRNA confers a gene-regulating functionality that is independent of the sequence of the RNA transcript; or (iii) regulation in *cis* depends solely on DNA elements within the lncRNA promoter or gene locus and is completely independent of the encoded RNA or its production (Kopp and Mendell, 2018).

Higher-order chromatin structure

Mammalian genomes encode genetic information in their linear sequence, but appropriate expression of their genes requires chromosomes to fold into complex three-dimensional structures (Gibcus and Dekker, 2013). The human genome contains approximately three billion bases equating to almost two metres of DNA in each nucleus with an average diameter of only 6 μm (Bickmore, 2013). Such packaging requires complex and dynamic spatial organisation to allow selective access to actively transcribed regions, compact unnecessary regions, and, during cell division, the ordered unravelling of chromatin (Dixon et al., 2012; Fraser et al.,

2015). This complex genomic folding necessary for physiological function follows a structural hierarchy described below from lowest to highest resolution (**Figure 1.3**).

At the largest scale, chromosomes occupy distinct nuclear territories. Gene-poor, mid-to-late-replicating chromatin is enriched in nuclear compartments located at the nuclear periphery and in perinucleolar regions (Cremer and Cremer, 2001). Individual chromosomes are then folded into A (open/active) and B (closed/silent) compartments that preferentially interact (Lieberman-Aiden et al., 2009). Within these compartments, the chromatin forms discrete megabase-scale topologically-associated domains (TADs) and smaller sub-TADs that correlate with genomic regions constraining chromatin spread (Dixon et al., 2012). At the kilobase-to-megabase scale, chromatin forms loop interactions that facilitate lineage-specific differential gene expression in a *cis*-regulatory manner by facilitating co-localisation of distal regulatory elements with gene promoters that span large genomic distances, even on different chromosomes (Hadjur et al., 2009; Schoenfelder et al., 2010). In turn, these act as transcriptional insulators by marking boundaries between different chromatin loops and domains (Ghirlando and Felsenfeld, 2016). TADs are largely conserved between cell types and across species, whereas sub-TAD topology varies in a tissue-specific manner, indicating that they are an inherent property of mammalian genomes (Phillips-Cremins et al., 2013; Vietri Rudan et al., 2015).

An important characteristic of these hierarchical DNA structures is the boundary elements that define them. In this respect, CCCTC-binding factor (CTCF) binds directly to a specific DNA motif and recruits the ring-shaped cohesin protein complex consisting of SMC1, SMC3, and RAD21 (Dixon et al., 2016; Rao et al., 2014). Therefore, in defining these boundaries, CTCF represents a fundamental epigenetic control element.

1.2 CTCF is a multi-functional DNA-binding protein

CTCF is a multifunctional eleven zinc finger nuclear phosphoprotein encoded by the *CTCF* gene. Combinatorial use of zinc finger domains allows protein binding to a range of specific ~50-60 bp DNA target sequences and proteins. Depending on the site of chromatin binding (and other co-bound proteins), CTCF may act as an activator or repressor of transcription, as described above. Additionally, CTCF forms methylation-sensitive insulators and regulates X-chromosome inactivation (Ohlsson et al., 2001). CTCF is, therefore, highly context- and cell type-specific, as described below.

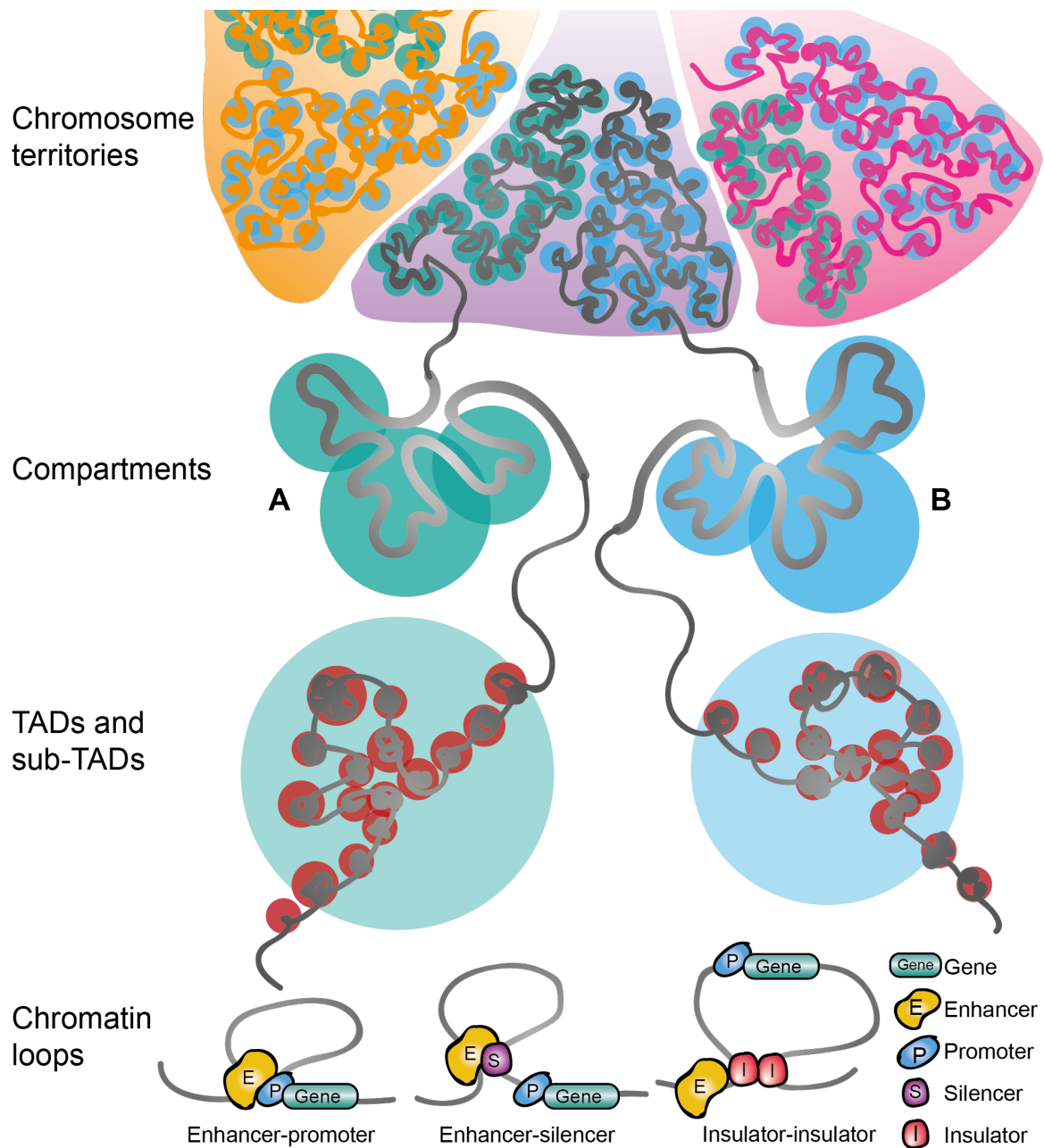


Fig. 1.3 Higher-order chromatin structure. Genomic scales of chromatin organisation are illustrated as a single unravelled chromatin fibre, from low (top) to high (bottom) spatial resolution. DNA occupies distinct nuclear territories (orange, purple, pink). DNA is broadly divided into euchromatic and heterochromatic compartments, here termed A (green) and B (blue), respectively. Self-interacting, megabase-scale topologically associated domains (TADs) and smaller chromatin loops further organise chromatin into regulatory units defined by binding of CTCF and cohesin complexes. Three examples of chromatin looping are shown: enhancer-promoter, enhancer-silencer, and insulator-insulator (Fraser et al., 2015).

1.2.1 CTCF in development, imprinting, and X inactivation

CTCF is necessary for embryonic development, and deletion of both *Ctcf* alleles in mice is embryonic lethal (Fedoriw et al., 2004; Moore et al., 2012). CTCF appears to play major roles in gene imprinting and X inactivation, which are important for normal development, gene silencing, and gene dosage control.

In mammals, almost all autosomal genes are expressed simultaneously by both alleles. However, ~1% of genes are imprinted; that is, one of the two inherited alleles is silenced, with reciprocal expression of either the maternally or paternally inherited allele (Wilkinson et al., 2007). CTCF plays an essential role in imprinting, which is an essential epigenetic mechanism in mammals for normal foetal growth and development (McGrath and Solter, 1984; Surani et al., 1984). Aberrant imprinting disturbs development and causes various inherited diseases (Reik and Walter, 2001) such as Prader-Willi syndrome (Nicholls et al., 1998).

The *H19* gene encodes one of the most abundant RNAs in the developing embryo and has served as a useful model for investigating mammalian imprinting. *H19* overexpression in transgenic mice results in late prenatal lethality, leading to the conclusion that dosage of its gene product is strictly controlled. RNase protection assays have been used to demonstrate that *H19* is paternally imprinted and only the maternal copy is expressed (Bartolomei et al., 1991). *H19* shares its enhancer elements with the maternally imprinted insulin-like growth factor 2 (*IGF2*) gene, and thus only the paternal copy is expressed (Reik and Walter, 2001). *IGF2/H19* locus imprinting is well characterised, and CTCF regulates parent-of-origin-specific monoallelic expression of *IGF2* and *H19* in a methylation-sensitive manner (Bell and Felsenfeld, 2000). A differentially methylated domain (DMD) upstream of *H19* contains CTCF binding sites: it is unmethylated on the maternal chromosome and, therefore, bound by CTCF, thus insulating the *IGF2* promoter from its enhancers. However, this DMD is methylated on the paternal allele, blocking CTCF binding and allowing downstream enhancers to interact with the paternal *IGF2* promoter (Lewis and Murrell, 2004). Under homeostatic conditions, the DMD remains hypomethylated during oogenesis, and RNA interference (RNAi) experiments in transgenic mice have shown that CTCF protects the *H19* DMD from *de novo* methylation (Fedoriw et al., 2004). Mutations in the CTCF binding region of the DMD of *H19* results in loss of *IGF2* imprinting and complex patterns of *de novo* methylation upon maternal inheritance (Pant et al., 2004), highlighting the importance of CTCF in the control of imprinting.

X chromosome inactivation is a dosage compensation mechanism in females that equalises gene expression by silencing one of the X chromosomes. In placental mammals, the chromosome selected for silencing is random but, after inactivation, the X chromosome of all descendants of that cell are inactivated in the same way. There is homologous pairing of the two X chromosomes during early embryonic development, which requires the presence of *trans*-acting CTCF (Xu et al., 2007). The lncRNA X-inactive specific transcript (*Xist*) then recruits specific chromatin modifiers, which results in the silenced X chromosome becoming a heterochromatic Barr body (Yang et al., 2015). However, a minority of genes on the inactive X chromosome, including *Xist*, are not silenced (Wutz, 2011) and instead are insulated by CTCF-constrained chromosomal loops (Filippova et al., 2005). This process is facilitated by lncRNAs including *Tsix* (the antisense transcript of *Xist*), which establishes an epigenetic switch for X inactivation (Chao et al., 2002; Filippova et al., 2005).

1.2.2 Transcriptional regulation

In addition to its developmental role, CTCF is essential in somatic cells, where it is integral to homeostatic gene expression. CTCF was initially characterised as a negative regulator of the proto-oncogene *c-MYC*, which plays a critical role in normal growth control, differentiation, and apoptosis (Evan and Littlewood, 1993) and is one of the most frequently affected genes in human cancers. CTCF binds ~200 bp upstream of the transcription start site (TSS) of *c-MYC*, which includes three CCCTC core sequence repeats and was proven experimentally by deleting 110 bp of the CTCF binding site in chicken embryonic fibroblasts, which resulted in a four- to eight-fold increase in *c-MYC* transcription (Lobanenkov et al., 1990). Thus, CTCF is a direct negative transcriptional regulator of *c-MYC*.

CTCF has exquisite evolutionary conservation over approximately 300 million years (Filippova et al., 1996), including in *Xenopus laevis* (African clawed frog) (Burke et al., 2002), *Danio rerio* (zebrafish) (Pugacheva et al., 2006), and *Drosophila melanogaster* (Moon et al., 2005). Although the exon-intron gene organisation of *CTCF* is different in mammals and birds, the open reading frame remains unchanged (Klenova et al., 1998). CTCF binding sites are also highly conserved across multiple mammalian species (Schmidt et al., 2012). This is in contrast to tissue-specific transcription factors such as hepatocyte nuclear factor 4 α (HNF4A) (Odom et al., 2007; Schmidt et al., 2010b; Stefflova et al., 2013) and enhancers (Villar et al.,

2015), which evolve much more rapidly. The sequences of some homologous CTCF DNA-binding regions, such as the *c-MYC* promoter, diverge between species: in humans, the *c-MYC* promoter it is bound by zinc fingers 3-11, whereas in chickens it is bound by zinc fingers 2-7 (Filippova et al., 1996). This divergence highlights the plasticity of zinc finger binding in order to maintain CTCF function in different species.

More recently, CTCF has been shown to regulate gene expression via long-range chromatin looping. A well characterised example of this phenomenon is at the locus control region (LCR) of the β -globin cluster, which interacts via long-range chromatin contacts to target genes in erythroid cells but not in those of other lineages such as neurons (Palstra et al., 2003; Splinter et al., 2006). The α -globin locus is also thoroughly characterised: paired convergent CTCF sites flank the α -globin regulatory region but only interact during erythropoiesis, thus defining a self-interacting erythroid compartment (Hanssen et al., 2017).

1.2.3 CTCF-dependent chromatin looping

The hierarchical structure of chromatin is necessary for genome function and proper gene regulation, and CTCF plays a critical role. However, a paradox exists with regard to the structure versus function of chromatin organisation, in that our knowledge of the global, low-resolution structure of chromosomal arrangement within the nucleus is very detailed, whereas our knowledge of chromatin function is more advanced at the finer, intragenic scale (Ruiz-Velasco and Zaugg, 2017). Three-dimensional chromatin organisation can be analysed using chromosome conformation capture technology, notably Hi-C, which allows genome-wide associations to be scrutinised (Dixon et al., 2012). The resolution of this technology has rapidly improved from megabase (Lieberman-Aiden et al., 2009) to kilobase resolution (Rao et al., 2014), and the dynamics of chromosomal organisation can now be evaluated in small cell numbers (Blanco et al., 2018; Hug et al., 2017) or even at single-cell resolution (Nagano et al., 2015).

As noted above, the boundaries of TADs and loops are enriched for CTCF binding sites (one at each end of the loop), where CTCF directly binds to specific DNA motifs to form a dimer (Dixon et al., 2012; Sofueva et al., 2013). The DNA motif is non-palindromic, and CTCF dimer formation occurs when the two CTCF binding sites are directionally convergent (de Wit et al., 2015; Guo et al., 2015; Rao et al., 2014), whereas inverted or disengaged CTCF sites do not necessarily

form chromatin loops (de Wit et al., 2015). Cohesin complexes are recruited to these boundary domains via the interaction of CTCF's C-terminus and the SA2 cohesin subunit (Xiao et al., 2011), where they help to establish and/or maintain topological domains, although CTCF and cohesin differentially contribute to chromatin organisation and gene regulation. For example, cohesin depletion causes a general loss of local chromatin interactions but the topological domains remain intact, whereas CTCF depletion reduces intradomain interactions but, notably, increases interdomain interactions (Zuin et al., 2014). Furthermore, depletion of cohesin or CTCF dysregulates distinct groups of genes (Sofueva et al., 2013; Zuin et al., 2014). Two independent mechanisms have been proposed to explain cohesin's control of chromatin organisation: (i) cohesin-independent segregation of the genome into fine-scale compartments defined by local transcriptional activity and the epigenetic landscape; and (ii) cohesin-dependent formation of TADs, which facilitates enhancer-promoter interactions (Schwarzer et al., 2017). The latter mechanism is debated but is currently hypothesised to occur either via the "handcuff model" (Vietri Rudan and Hadjur, 2015) or the "loop extrusion" model (Sanborn et al., 2015) (**Figure 1.4**).

The handcuff model proposes that CTCF serves as the initial, static binding factor that defines a grid of potential insulation sites based on high-specificity sequence motifs. Cohesin complexes are then recruited to the CTCF grid to engage in preferential interactions that give rise to long-range chromosomal loops, which effectively have an insulatory effect and thus organise chromosomes into domains (Vietri Rudan and Hadjur, 2015). This is supported by the known co-localisation of CTCF and cohesin at TAD boundaries and that removal of CTCF and/or cohesin alters chromatin architecture. However, given that there are far more CTCF binding sites than TADs, it is unclear which, when, and why two sites interact.

The loop extrusion model is more dynamic and proposes that an "extrusion complex" containing two tethered DNA-binding subunits (likely including CTCF-cohesin heterodimers) is loaded onto chromatin, with both DNA-binding subunits binding in close spatial proximity to form a tiny chromatin loop. The binding units are thought to travel along the chromatin fibre in opposite directions until they reach a pair convergent CTCF binding motifs, while DNA extrudes through the cohesin loops (Dekker and Mirny, 2016; Sanborn et al., 2015). This model is increasingly favoured, especially since convergent CTCF sites are a feature of TAD boundaries (de Wit et al., 2015; Guo et al., 2015; Rao et al., 2014), and genome-editing experiments to delete (Sanborn et al., 2015) or invert (de Wit et al., 2015) CTCF binding sites result in alterations rather than collapse of TADs, which could be computationally predicted

based on known CTCF sites. However, this model raises other questions: what are the "extrusion complex" proteins; how does the energy expenditure required for such machinery favour the cell; can the same machinery work in euchromatic and heterochromatic compartments; how is the process affected by (or impact) transcription and replication; and how does this account for different levels of chromatin structure (i.e., loops/sub-TADs within TADs) (Dixon et al., 2016)?

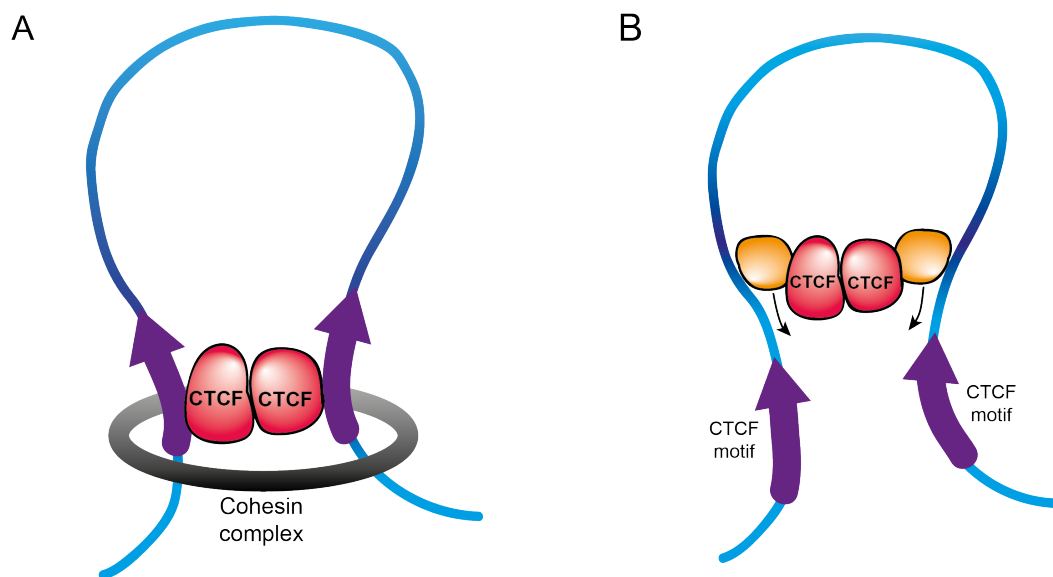


Fig. 1.4 Models of TAD formation. (A) The "handcuff" model. TADs are formed by CTCF (red) and cohesin complexes (grey) connecting two CTCF boundary sequences (purple). (B) The "loop extrusion" model. A pair of tethered CTCF proteins bound to loop extrusion factors (orange) extrude the chromatin fibre; two CTCF molecules slide along the chromatin fibre in opposite directions before pausing at converging CTCF DNA binding motifs (Dixon et al., 2016).

1.3 Cancer epigenetics

Given the fundamental importance of epigenetics in maintaining DNA, cellular, and organism homeostasis while simultaneously facilitating appropriate responses to developmental or environmental cues, it is perhaps unsurprising that epigenetic dysfunction has been increasingly scrutinised in disease pathogenesis and that epigenetic defects have been found to contribute to cancer initiation and development (Berger et al., 2009; Egger et al., 2004; You and Jones, 2012).

Our >20,000 protein-coding genes (Genome Reference Consortium Human Build 38 (GRCh38)) account for less than 2% of the genome. Epigenetic features of the non-protein-coding genome such as DNA methylation, histone marks, noncoding RNAs, and chromatin structure must, therefore, be integral to cellular homeostasis (Alberts et al., 2014). The proportion of a species' genome that is non-protein-coding is thought to positively correlate with its biological complexity (Liu et al., 2013a).

The importance of the non-protein-coding genome in humans is further highlighted by human cancer analyses, which show that >80% of mutations that are significantly associated with disease phenotypes arise in non-protein-coding regions (Welter et al., 2014). Such genome-wide association studies (GWAS) have revealed that somatic mutation rates vary between tumour types (Alexandrov et al., 2013) and frequency varies across the genome (Lawrence et al., 2013). Mutation rate is also associated with transcriptional activity (Chapman et al., 2011), chromatin state (Schuster-Böckler and Lehner, 2012), replication timing (Liu et al., 2013b; Stamatoyannopoulos et al., 2009), and nuclear topology (Smith et al., 2017). Furthermore, regulatory sites show elevated substitution rates in cancers (Kaiser et al., 2016), most notably CTCF/cohesin binding sites (Katainen et al., 2015; Sabarinathan et al., 2016). The underlying mechanisms are not well understood (Flavahan et al., 2017; Stricker et al., 2016). The high mutational burden observed in protein-bound DNA regions of melanomas and lung tumours is caused by impaired nucleotide excision repair (NER) (Sabarinathan et al., 2016), and the reduced exonic mutation rate is due to differential mismatch repair (MMR) rather than purifying selection (Frigola et al., 2017). Such findings have opened up new fields of human disease research including epigenome-wide association studies (EWAS) (Lappalainen and Grealis, 2017; Michels et al., 2013).

Epigenetic diseases may be inherited or somatic: Angelman's syndrome, Prader-Willi syndrome (Nicholls et al., 1998), and Beckwith-Wiedemann syndrome (Maher and Reik, 2000) arise due to faulty imprinting, while many cancers are known to harbour somatic epigenetic defects (Sandoval and Esteller, 2012). Global changes in the epigenetic landscape are considered a hallmark of ageing (López-Otín et al., 2013) and cancer (Sandoval and Esteller, 2012) since histone modifications, DNA methylation changes, and altered expression of chromatin-modifying enzymes result in global changes in gene expression and the development and/or progression of disease (Hanahan and Weinberg, 2011; Sharma et al., 2010).

The first description of the contribution of epigenetics to carcinogenesis was in colorectal cancer (CRC). These tumours often harbour inactivating somatic mutations in

the MMR gene human MutL Homologue 1 (*hMLH1*). However, a proportion of CRCs were found to not express *hMLH1* despite a normal coding sequence. Instead, all these cases, in contrast to *hMLH1*-mutated tumours, exhibited cytosine methylation of the *hMLH1* promoter region (Kane et al., 1997). This epigenetic silencing was quickly recognised as an alternative loss-of-function pathway satisfying Knudson's two-hit hypothesis (Jones and Laird, 1999), and *de novo* promoter hypermethylation of tumour suppressor genes is now a well-recognised mechanism contributing to cancer initiation and progression (Jones and Baylin, 2002).

It has since emerged that chromatin and epigenetic aberrations have the potential to confer on cells the full range of oncogenic properties represented in the classic "hallmarks" depiction of cancer (Hanahan and Weinberg, 2000, 2011). Genetic, environmental, and metabolic factors can make chromatin aberrantly permissive or restrictive. Restrictive chromatin states may prevent appropriate tumour suppressor functions or block differentiation, whereas permissive chromatin creates a state of "epigenetic plasticity" that can activate oncogene expression or cell fate changes that drive carcinogenesis (Flavahan et al., 2017). As for coding mutations, many epigenetic aberrations will be stochastic, inconsequential "passengers", whereas a minority that confer a fitness advantage to a cell will be selected as "drivers" (Flavahan et al., 2017). For example, specific point mutations in the gene encoding histone protein H3.3 result in a global reduction of the repressive histone mark H3K27me3 and DNA hypomethylation. This has been shown to activate gene expression and promote tumourigenesis in a clinically distinct group of high-grade gliomas (Bender et al., 2013). Such brain tumours exemplify the interplay between genetic and epigenetic mechanisms in tumourigenesis, with varying contributions of each in glioblastoma (a brain tumour that primarily affects adults), ependymomas (predominantly a paediatric tumour), and anaplastic astrocytomas. As mentioned above, in glioblastoma, most hallmarks can be traced to genetic drivers, whereas ependymomas are dominated by epigenetic factors including DNA hypermethylation but lack recurrent mutations (Lathia et al., 2015; Mack et al., 2014). Anaplastic astrocytomas are often driven by both genetic and epigenetic lesions, leading to different hallmarks (Flavahan et al., 2017). As well as the classical epigenetic events described above, ncRNAs are also implicated in carcinogenesis. For example, the miRNAs miR-15 and miR-16, which usually silence the anti-apoptotic *BCL2* gene, are downregulated in chronic lymphocytic leukaemia (Ventura and Jacks, 2009).

Finally, systematic enrichment analysis of prostate cancer GWAS hits have been used to attempt to decipher the underlying molecular mechanisms of cancer. In

addition to mutations within exonic regions, well-characterised transcription factor binding sites of androgen receptor (AR), FOXA1, ETS-related gene (ERG), and CTCF binding regions were significantly enriched for prostate cancer risk single nucleotide polymorphisms (SNPs) (Chen et al., 2015). This study suggests that although most cancers occur as a result of multiple regulatory aberrations, CTCF is an important cancer protein.

1.3.1 *CTCF* is a tumour suppressor gene

There is mounting evidence that, in addition to its functions described above, CTCF is involved in the pathogenesis of cancer (Ohlsson et al., 2001). However, its mechanism of action in the initiation or progression of carcinogenesis is poorly understood.

The human *CTCF* gene lies in cytogenetic band 16q22.1 on chromosome 16, a region found to be frequently deleted in sporadic breast and prostate cancers over twenty years ago (Filippova et al., 1998). *CTCF* was, therefore, proposed as a candidate tumour suppressor gene. Furthermore, microdeletions of CTCF binding sites on chromosome 11p15 in patients with Beckwith-Wiedemann syndrome leads to a 600-fold increased risk of Wilms' tumour, a kidney cancer that contains malignant cells of different embryological origin (Prawitt et al., 2005). More recently, human genome sequencing studies have identified *CTCF* as a putative driver gene in several human cancers as detailed below, in keeping with the action of a tumour suppressor gene.

Several studies have been performed that have attempted to characterise CTCF's mechanism of tumour suppression. However, these studies have largely considered how loss of CTCF function (Recillas-Targa et al., 2006) gives rise to cancer hallmarks such as sustained proliferation, invasion and metastasis, replicative immortality, genomic instability, and mutations via single gene effects (Hanahan and Weinberg, 2000, 2011). Since CTCF is a global regulator of gene expression, altered CTCF expression levels and DNA binding affinity are likely to have widespread downstream effects on the transcriptome and proteome, and these global effects, particularly from the epigenomic perspective, are poorly characterised.

CTCF is a key transcriptional regulator of the tumour suppressor gene *TP53* (Saldaña-Meyer and Recillas-Targa, 2011) via physical interaction with its anti-sense RNA *WRAP53* and miRNAs (Saldaña-Meyer et al., 2014). Mutant *CTCF* results in reduced *TP53* expression (Saldaña-Meyer and Recillas-Targa, 2011) and

defective TP53 responses to DNA damage (Saldaña-Meyer et al., 2014), which, if left uncorrected, leaves the cell susceptible to further mutagenic insults. CTCF is also responsible for silencing oncogenes such as *BCL6*: CTCF-mediated silencing of *BCL6* is inhibited by hypermethylation of the intronic DNA-binding site in B cell lymphoma cell lines (Lai et al., 2010).

Appropriate expression of pro-apoptotic factors is essential for maintenance of normal cell turnover and tissue homeostasis. Apoptosis-associated tyrosine kinase (AATK), which acts as a tumour suppressor gene, is epigenetically activated by CTCF. Promoter hypermethylation silences the gene and is thought to be an early event in carcinogenesis. The AATK promoter is frequently hypermethylated in human tumours and cell lines, resulting in deficient CTCF binding and reduced AATK expression (Haag et al., 2014). As a result, there is failure of growth inhibition, migration, and apoptosis (Ma and Rubin, 2014).

As detailed above, higher-order chromatin structure is also essential for co-ordinated gene expression, with CTCF playing an integral role in this process. CTCF/cohesin complexes selectively bind to unmethylated DNA, and it has been shown that altered methylation influences CTCF-dependent chromatin looping and alters downstream gene expression (Kang et al., 2015). Given that *de novo* methylation is an epigenetic hallmark of cancer, its effect on CTCF binding and chromatin loop formation (Kang et al., 2015) provides a possible carcinogenic mechanism.

Large-scale chromosomal rearrangements, which can result in gene fusions, are well described in cancer (Rabbitts, 1994, 2009). More recently, the discovery of RNA *trans*- and *cis*-splicing between adjacent genes (cis-SAGe) has provided another mechanism by which fusion RNAs that contain exons belonging to neighbouring genes from the same strand can be generated. The role of CTCF in generating these fusion RNAs has been investigated in prostate cancer, where CTCF silencing resulted in cis-SAGe events, usually resulting in the production of noncoding fusion RNAs (Qin et al., 2015). This study identified a range of fusion RNAs in cell lines, human prostate cancers, and control samples. Although CTCF was not explicitly causative in creating fusion RNAs in cancer, *Ctcf* silencing did change the resultant gene fusions, and the possibility of CTCF-induced fusion proteins remains a possibility.

However, the most convincing experimental evidence that *CTCF* may act as a tumour suppressor is from Kemp et al. (2014), who used an *in vivo* mouse model of *Ctcf* haploinsufficiency to demonstrate a multi-lineage predisposition to spontaneous

carcinogenesis. This study, which provided *in vivo* rationale for the main hypotheses tested in this body of work, is described in detail below.

1.3.2 *Ctcf* haploinsufficiency predisposes to cancer in mice

Using a germline *Ctcf* haploinsufficient mouse model, Kemp et al. (2014) found that *Ctcf*^{+/-} mice were more susceptible to spontaneous, radiation-, and chemically-induced cancers compared to C57BL6/129 wild-type littermates. Haploinsufficient mice developed tumours earlier (80% of *Ctcf*^{+/-} vs. 40% of wild-type littermate mice were euthanised due to cancer by 100 weeks) and were three-times more likely to develop multiple tumours. Furthermore, the tumours were histologically more aggressive, and metastatic disease was more frequent. They concluded that hemizygous loss of *Ctcf* enhances malignant progression. They also used urethane to induce pulmonary neoplasms in *Ctcf*^{+/-} and wild-type mice: *Ctcf*^{+/-} mice developed more frequent and significantly larger lung tumours than wild-type mice, which resulted in earlier death in *Ctcf*^{+/-} mice. Furthermore, the type of neoplasms varied between these groups: 77% of lung tumours from wild-type mice were benign adenomas, whereas 69% of *Ctcf*^{+/-} lung tumours were malignant adenocarcinomas, suggesting that CTCF is important in both cancer initiation and progression.

Methylation analysis of non-neoplastic tissue showed that differentially methylated sites clustered according to genotype, suggesting that *Ctcf* hemizygosity results in altered DNA methylation. Methylation was altered at the *Igf2/H19* imprinting region; however, the tumour-suppressor genes *p16*, *p19*, and *Mlh1* were not altered and overall methylation patterns were stable across genotypes. This implies that *Ctcf* only regulates DNA methylation at very specific loci and that methylation alone does not fully account for the predisposition to cancer seen in these mice. Since these *Ctcf* mice were germline haploinsufficient, the authors could not delineate the precise mechanism of cancer susceptibility due to potential global effects on metabolism and immunity.

1.3.3 Mutation of *CTCF* and its binding sites in human cancers

In addition to *in vitro* and *in vivo* studies, *CTCF* mutations have been shown to be important in human cancers. Frequent *CTCF* mutations (single base substitutions, deletions, and insertions) have been identified in human malignancies including uterine, breast, liver, oesophageal, head and neck, and pancreatic cancers and

lymphoma (ICGC Data Portal, <https://dcc.icgc.org/genes/ENSG00000102974>). In a study of 4,623 whole exome sequences from 13 cancer sites, *CTCF* was identified as a putative driver gene in uterine endometrioid carcinoma, head and neck squamous cell carcinoma, and breast carcinoma. All mutations were protein-coding, loss of function mutations resulting in missense coding alterations or truncation of the transcript (Gonzalez-Perez et al., 2013; Rubio-Perez et al., 2015), in keeping with the action of tumour suppressor gene. Lawrence et al. (2014) analysed 4,742 human cancers of 21 histological types for somatic point mutations and identified *CTCF* and only 21 other genes as significantly associated with three or more histological tumour types. In addition to *CTCF*, this geneset included well-characterised cancer drivers including *TP53*, *PIK3CA*, *PTEN*, *RB1*, *KRAS*, *NRAS*, and *BRAF*. These data highlight the important biological role for *CTCF* in different tissue types and implicate it as a tumour suppressor gene in humans.

Altered CTCF expression levels have also been documented in human cancers. CTCF expression was measured in 71 pre-invasive and invasive cervical neoplasms and found to be significantly overexpressed in invasive carcinomas compared to pre-invasive lesions or normal controls (Velázquez-Hernández et al., 2015). However, this study was not functionally validated and no mechanism of action was proposed. Data from COSMIC showed CTCF expression to be up- and downregulated with almost equal frequency across numerous cancer tissue types (Forbes et al., 2017). In keeping with this, low CTCF expression was correlated with favourable survival outcomes in liver cancer but poorer outcomes in renal cancer (Uhlen et al., 2015), thus illustrating a degree of pleiotropism in CTCF's action in human cancers.

More recent studies have shown that, in addition to mutations in *CTCF* itself, mutations in CTCF DNA-binding sites are common in cancer (Katainen et al., 2015). Regulatory regions contain mutations under selective pressure, suggesting a greater role for regulatory mutations in cancer than previously recognised (Kaiser et al., 2016; Melton et al., 2015). Computational analysis of 1,073 published cancer genomes across eleven tissue types (Hudson et al., 2010) showed that multiple cancer types accumulate point mutations at CTCF/cohesin-binding sites, highlighting that mutational hotspots arise in the non-protein-coding as well as coding cancer genome (Katainen et al., 2015). The increased rate of mutations at CTCF binding sites was especially prominent in melanoma, with the accumulation of functional mutations at CTCF/cohesin-binding sites proposed to be due to uneven nucleotide excision repair across the motif rather than an inherent susceptibility to mutation (Poulos et al., 2016). It has also been suggested that this repair process is physically

impaired by the binding of transcription factors at the motif (Sabarinathan et al., 2016).

Examples of CTCF binding site mutations can be found in other tumour types. Gain-of-function *IDH1* mutations are initiating events that define major clinical and prognostic classes of gliomas due to the production of the onco-metabolite 2-hydroxyglutarate by the mutant IDH1 protein. Human *IDH1*-mutant gliomas have specific hypermethylation at CTCF binding sites, compromising CTCF binding and in turn causing loss of insulation between topological domains and aberrant gene activation. In contrast, clustered regularly interspaced short palindromic repeat (CRISPR)-mediated disruption of the CTCF motif in *IDH1* wild-type gliomaspheres upregulated *PDGFRA* and increased proliferation, suggesting that *IDH1* mutations promote gliomagenesis by disrupting chromosomal topology and allowing aberrant regulatory interactions (Flavahan et al., 2015). In lung cancer, CTCF binding site variants were analysed in 2,331 lung cancers and 3,077 control samples and the rs60507107 SNP was identified as a novel lung cancer susceptibility locus, which was then validated in an independent cohort (Dai et al., 2015).

It is likely that many more noncoding driver mutations at CTCF binding sites and other regulatory sites will be discovered over the next few years. The Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium (Campbell et al., 2017) is expected to identify common mutational patterns in more than 2,800 cancer whole genomes (Sabarinathan et al., 2017), thereby exploring the nature and consequence of somatic and germline variations in both coding and noncoding regions. Given that there is specific emphasis on *cis*-regulatory sites, noncoding RNAs, and large-scale structural alterations, this study is likely to yield further insights into CTCF's role in cancer.

1.4 Mouse models of liver cancer

Hepatocellular carcinoma (HCC) is the predominant form of primary liver cancer and the second most common cause of cancer death worldwide (Torre et al., 2015). HCC develops in the context of chronic liver disease in 80% of cases, in which chronic hepatocyte injury leads to genetic damage that underpins HCC (Hardy and Mann, 2016). The incidence of HCC is increasing globally, but treatment options are limited and prognosis remains poor. The use of tractable experimental models is, therefore, essential in order to develop a more meaningful molecular classification of HCC, in turn offering the promise of predictive and prognostic biomarkers and potential

therapeutic targets (personalised or precision medicine), or to identify targets to prevent progression of liver disease from a benign to malignant state (Bakiri and Wagner, 2013).

An ideal model of liver cancer should: (i) reproduce key biological and cellular features (hepatocyte damage, degeneration, regeneration, proliferation, and transformation) in spontaneous models of HCC; (ii) recapitulate the molecular and cellular events in HCC development and progression; (iii) adequately reflect tumour/host interactions; (iv) allow the study of the primary tumour and metastases; (v) mimic the human tumour microenvironment; and (vi) be affordable and easy to manipulate (Li et al., 2012). There are numerous mouse models of liver cancer, including exposures to chemical carcinogens and genetic alterations that reproduce different aspects of the phenotypic, biological, and molecular events that occur during hepatocarcinogenesis. It is important to select the most appropriate model to address the specific scientific question of interest, and a few important considerations are detailed below (Newell et al., 2008).

Mouse strains

Inbred laboratory mice show marked variability in their susceptibility to spontaneous liver tumours: C3H mice are highly susceptible to hepatocarcinogenesis whereas C57BL/6 mice (Maronpot, 2009; Puccini et al., 2013) and wild-derived mouse strains tend to be very tumour resistant (Gu  net and Bonhomme, 2003). Such strain difference provide an opportunity to study spontaneous tumourigenesis in the absence of chemical induction and/or engineered genetic aberrations. However, this also means that the strain used is an important variable to consider when planning experiments since strain selection may affect the experimental results.

Chemical models

By far the most commonly used chemical carcinogenesis model uses intraperitoneal (IP) diethylnitrosamine (DEN) to induce liver tumours in rodents (Rajewsky et al., 1966). Bioactivation in centrilobular (zone 3) hepatocytes results in a genotoxic alkylating agent, which causes strain-dependent oncogenic mutations (Buchmann et al., 2008), dysplastic changes, and subsequent progression to HCC. If animals are treated after the proliferative post-natal stage (up to 15 days old), an additional potentiating stimulus is required such as partial hepatectomy or phenobarbital treatment (Aleksic et al., 2011).

Genetically-engineered mouse (GEM) models

A wide range of genetically-engineered mouse (GEM) models have been developed to study the breadth of alterations observed in human cancers. Transgenic mice can be engineered to express oncogenes or tumour suppressor genes in a non-physiological manner due to the introduction of ectopic promoter and enhancer elements. This is traditionally achieved by microinjection of recombinant DNA directly into the pronucleus of a fertilised mouse egg or, alternatively, by gene targeting ("knock-in") and lentiviral transduction in embryonic stem cells (Frese and Tuveson, 2007). Transgene expression can be reversibly and temporally controlled using a tetracycline (tet)-inducible system (Furth et al., 1994), in which the inducing agent reversibly activates the target gene of a chimeric transcriptional activator (Newell et al., 2008). Alternatively, conditional mouse models, such as the Cre-Lox system, allow both spatial and temporal control of target genes in specific cells, tissues, or organs. In this system, the target gene is flanked by 34 bp *loxP* sites, and when Cre recombinase (isolated from bacteriophage P1) is expressed there is site-specific recombination between the *LoxP* sites, resulting in excision of the target gene (Sauer and Henderson, 1988).

Since mice are resistant to infection by human hepatitis B (HBV) or C (HCV) viruses, the first transgenic mice were developed to model the chronic carrier state of HBV infection, in which viral DNA sequences integrate into the host genome (Babinet et al., 1985; Chisari et al., 1985). Together with similar models of HCV, these transgenic mouse models have provided experimental evidence that viral hepatitis genes can initiate and promote liver cancer independent of oncogenes or environmental carcinogens (Li et al., 2012). Gene targeting approaches have been used to induce a range of molecular events in hepatocarcinogenesis including in cell cycling pathways (p53, Rb, E2F, and SV40), telomeres, growth factor signalling pathways (TGF- α , c-Myc, hepatocyte growth factor, and c-Met pathway), PTEN/Akt/mTOR signalling pathway, IGF and EGF signalling pathway, and the Wnt/ β -catenin pathway (Newell et al., 2008).

More recently, CRISPR and CRISPR-associated proteins (Cas) have offered a powerful means to edit the genome (Jinek et al., 2012). CRISPR/Cas9 can be applied to mouse models to produce gene knockouts, deletions, point mutations, and short insertions such as *loxP* sites (Yang et al., 2013). CRISPR/Cas9 is generally cheaper and quicker than traditional, directed mutagenesis methods, and provides the additional option of being able to alter multiple genes simultaneously.

1.5 Thesis outline

CTCF is, as outlined above, a global regulator of gene expression, and altered CTCF expression levels and DNA binding affinity are likely to have widespread downstream effects on the transcriptome and proteome. These global effects, particularly from the epigenomic perspective, are poorly characterised. CTCF is also regarded as a putative tumour suppressor *in vitro* and *in vivo*, with germline *Ctcf* haploinsufficient mice rendered susceptible to spontaneous, radiation-, and chemically-induced tumours of multiple lineages. Although several plausible tumour suppressor mechanisms for CTCF are postulated, the exact mechanisms by which *CTCF* mutations contribute to cancer initiation and progression have not been fully established. This summary highlights two major knowledge gaps: first, the nature and extent of global CTCF-mediated changes on the transcriptome and proteome, and second, the mechanism by which *CTCF* mutations predispose to cancer. This set of studies aims to address these knowledge gaps.

We therefore hypothesised that:

1. **A reduced genomic concentration of *Ctcf* increases cancer susceptibility by altering chromatin homeostasis**; reducing CTCF concentrations would be expected to have a quantitative effect on DNA binding, resulting in downstream effects on transcription and protein expression dependent on their relationship with altered sites of CTCF activation or repression and/or altered chromatin CTCF-dependent looping.
2. **Liver tumourigenesis is accelerated in *Ctcf* conditional knockout mice** due to alterations in the expression of cancer-related pathways.

This study first uses functional genomics and quantitative proteomics in an *in vitro* model to examine the effect of a reduced genomic concentration of *Ctcf* on genome-wide CTCF binding and the downstream consequences for transcription and translation. In the second set of experiments, a conditional, hepatocyte-specific *Ctcf* hemizygous mouse is interrogated to establish the tissue-specific effects of *Ctcf* haploinsufficiency and its impact on hepatocarcinogenesis when challenged with a chemical carcinogen.

Chapter 2

Materials and methods

2.1 Mouse colony management

All animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical Review Body (AWERB). All animals were maintained using standard husbandry: mice were group housed in Tecniplast GM500 Mouse IVC Green Line cages in a room with a 12 hour light / 12 hour dark cycle and *ad libitum* access to water and food (LabDiet 5058). Cages contained aspen bedding and the following cage enrichments: nesting material, aspen chew stick, and cardboard tunnel.

The study was designed to be in keeping with the principles of the 3Rs: replacement, reduction, and refinement. Animal models cannot be replaced, since the nature of the study is to investigate *Ctcf* deficiency in whole organisms using tissue-specific analysis. Current 3D cell culture or organoid models are inadequate to answer this biological question. The experiments have been designed to reduce the number of animals required. Any excess tissue harvested was added to the biorepository for follow-up studies to ensure maximum utilisation of the biological material. Refinement was optimised, since the mice were subject to either no intervention or a single intra-peritoneal injection.

2.1.1 Genetically engineered mice

Cre-Lox recombination was used in the generation of both mouse strains. This system uses a single enzyme, Cre recombinase, to recombine a pair of short target sequences called *Lox* sequences, allowing deletion of a specific genomic locus.

***Ctcf* hemizygous mice**

The following mice were used to generate *Ctcf* hemizygous mice: male C57BL/6 mice carrying the modified *Ctcf* allele in which *loxP* sites are inserted at exon 3 and exon 12 (*Ctcf^{tm1.1Laat}/Ctcf^{tm1.1Laat}*) (Heath et al., 2008) (founder mouse was a gift from the Galjart lab, Erasmus MC, Rotterdam); and female B6.C^{Tg(Pgk1-cre)1Lni/Crs}/J mice (JAX stock #020811) (Lallemand et al., 1998), which express Cre recombinase under the control of the phosphoglycerate kinase 1 (*Pgk1*) promoter. When crossed with a strain containing a *loxP* site-flanked sequence, Cre-mediated recombination results in deletion of the flanked sequence. Cre activity commences at the diploid phase of genesis thus ensuring that all cells of the resulting hemizygous zygote contain the targeted *Ctcf* allele.

The resulting *Ctcf* hemizygous (*Ctcf^{+/-}*) line was maintained by back-crossing on a C57BL/6 background to produce *Ctcf* hemizygous (*Ctcf^{+/-}*) and wild-type (*Ctcf^{+/+}*) offspring. These mice were used to: (i) generate mouse embryonic fibroblast cultures and (ii) study the reported tumour susceptibility of these mice over prolonged ageing.

Hepatocyte-specific *Ctcf* hemizygous mice

The following mice were used to generate hepatocyte-specific *Ctcf* hemizygous mice: male *Ctcf^{+/-}* mice (generated above) and female B6.Cg-*Speer6-ps1^{Tg(Alb-cre)21Mgn}*/J mice (JAX stock #003574) (Postic et al., 1999). Albumin is highly and uniquely expressed by hepatocytes and therefore the "*Alb-cre*" mouse is very efficient for liver-specific gene knockouts using the Cre/*loxP* system; all non-hepatocytes in the mouse retain two copies of *Ctcf*.

2.1.2 Mouse genotyping

Genotyping of mouse ear biopsies or embryonic tissue was performed by polymerase chain reaction (PCR) using primers designed for: wild-type *Ctcf*, floxed *Ctcf*, excision of floxed *Ctcf*, and *Cre* (**Table 2.1**). Mice positive for both floxed-*Ctcf* and *Cre* are expected to have hemizygous loss of *Ctcf* in hepatocytes.

2.2 Mouse embryonic fibroblast cultures

Male *Ctcf^{+/-}* mice were crossed with C57BL/6J pro-oestrus or oestrus females that were then monitored daily for the presence of a vaginal plug. The date of plug

Table 2.1 Primers used for mouse genotyping

Probe	Forward primer	Reverse primer
<i>Ctcf</i> - wild-type	ATGGTTGTGAGCCACCATGTGA	AAGCACTGATTGCTCTAAAGAAGGTTGT
<i>Ctcf</i> - floxed	GCTGGGCTCGACTCTAGACATAT	GCTCTAAAGAAGGTTGTGAGTTCTGA
<i>Ctcf</i> - excised	GCTGGGCTCGACTCTAGACATAT	GCAAACCTCCATCTCTAGCCTCTCTA
<i>Cre</i>	TTAATCCATATTGGCAGAACGAAAACG	CAGGCTAAGTGCCTTCTCTACA

identification was considered embryonic day 0.5 (E0.5). Mice were euthanised by cervical dislocation on day E13.5.

Using sterile technique, both uterine horns containing embryos were removed from the mouse and placed in ice cold sterile phosphate-buffered saline (PBS). Each embryo was processed separately in order to eliminate the risk of contamination with tissue of a different genotype. The amniotic sac and visceral organs were dissected and discarded; the head was removed and used for genotyping with real-time PCR. The remaining embryonic tissue was minced and trypsinised at 37°C for 30 min, quenched with “MEF media” (Dulbecco’s Modified Eagle Medium (DMEM, Gibco; Thermo Fisher Scientific, Waltham, MA) supplemented with 4.5 g/L D-glucose, L-glutamine and pyruvate, 10% heat-inactivated foetal bovine serum (FBS, Gibco), 1% amphotericin B antimycotic (Life Technologies, Carlsbad, CA), and 1% penicillin-streptomycin solution), and each embryo suspension was seeded into a 15 cm dish and incubated at 37°C in 5% CO₂. The media was refreshed after 24 h. When confluent, cultures were split 1:3 in the absence of antibiotics from passage 2 onwards.

ChIP-seq, RNA-seq, proteomic, and Hi-C experiments were all performed from a single passage 4 (P4) culture. Mouse embryonic fibroblast (MEF) cultures for each biological replicate were expanded and harvested in pairs, one wild-type and one *Ctcf* hemizygous line at a time, to control for culture-related batch effects.

2.3 qPCR

Total RNA was extracted from P4 MEF cultures from six biological replicates from each genotype using QIAzol Lysis Reagent (Qiagen, Hilden, Germany), according to the manufacturer’s instructions. cDNA was synthesised from RNA using the High-Capacity RNA-to-cDNA Kit (Thermo Fisher Scientific), and qPCR was performed in three technical replicates using TaqMan probes (Thermo Fisher Scientific) according

to the manufacturer's instructions. *Ctcf* mRNA levels were calculated using mean Ct values normalised to *Gapdh* signal for each pair of MEF cultures.

2.4 Quantitative western blotting

Protein was extracted from P4 MEFs from six biological replicates from each genotype: cells were washed with ice cold PBS, lysed in radioimmunoprecipitation assay (RIPA) buffer (50 mM Hepes-KOH pH 7.6, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Na-deoxycholate), pulse sonicated on ice (3 x 10 s), agitated for 30 min at 4 °C, the cell debris was pelleted by centrifugation, and the supernatant was quantified using Direct Detect Infrared Spectrometer (Merck Millipore, Burlington, MA). 20 µg total protein was run on a 4-12% Bis-Tris gel and transferred to a membrane using an iBlot 2 Gel Transfer Device (Thermo Fisher Scientific). The membrane was blocked using Odyssey Blocking Buffer in TBS (LI-COR Biosciences, Lincoln, NE), incubated overnight with CTCF anti-rabbit antibody (Cell Signaling Technology, Danvers, MA; D31H2 XP) (1:1000) and β -actin anti-mouse antibody (Sigma-Aldrich, St. Louis, MO; clone AC-74) (1:5000). The membrane was washed 4 x 5 min in TBS + 0.1% Tween and incubated for 45 min at room temperature (RT) with fluorescent-conjugated infra-red (LI-COR Odyssey) antibodies: goat anti-mouse antibody (1:20,000) labelled with 680 LT infrared dye (P/N 925-68070) and goat anti-rabbit antibody (1:5000) with 800 CW infrared dye (P/N 925-32211). The membrane was washed a further four times before visualisation and quantification using the Odyssey CLx Imaging System. Relative CTCF abundance was calculated for each pair of MEF cultures using normalised fluorescence values using β -actin as the loading control.

2.5 Chromatin immunoprecipitation followed by high-throughput sequencing

Formaldehyde cross-linking

MEFs for six biological replicates from each genotype were fixed in DMEM containing 1% fresh formaldehyde and incubated at RT for 10 min, quenched with 250 mM glycine for 10 min, and washed twice with ice cold PBS. The fixed cells were lifted off the plate, pelleted by centrifugation, washed in PBS containing 1x Complete

Protease Inhibitor (PI) Cocktail (Roche, Welwyn Garden City, UK), and flash-frozen at -80 °C.

Pre-block and antibody binding to magnetic beads

100 μ l protein G magnetic Dynabeads (Life Technologies) were washed three times with 1 mL blocking solution (0.5% bovine serum albumin (BSA) (w/v) in PBS) then re-suspended in 250 μ l blocking solution with 10 μ g of antibody and incubated overnight at 4 °C on a rotating platform. The following antibodies were used: CTCF (rabbit polyclonal, Merck Millipore 07-729, lot 2517762); H3K4me3 (mouse monoclonal IgG clone CMA304, Merck Millipore 05-1339, lot 2603814); H3K27ac (rabbit polyclonal IgG, Abcam (Cambridge, UK) 4729, lot GR244014-1); Rad21 (rabbit polyclonal, Abcam 922, lot GR12688-9). The beads were washed a further three times the following day immediately prior to adding them to the sonicated chromatin.

Nuclear lysis

Each cell pellet was suspended in 1 ml Lysis Buffer 1 (LB1 (50 mM Hepes-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% nonyl phenoxypolyethoxylethanol (NP-40), 0.25% Triton X-100)) containing 1 x PI, incubated at 4 °C for 10 min, and centrifuged at 2000 x g for 5 min at 4 °C. The pellet was re-suspended in 1 ml Lysis Buffer 2 (LB2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA)) containing 1 x PI, incubated at 4 °C for 5 min, and centrifuged again at 2000 x g for 5 min at 4 °C. Finally, the pellet was re-suspended in 300 μ l Lysis Buffer 3 (LB3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine)) containing 1 x PI. Cells from four 15 cm dishes were pooled for each chromatin immunoprecipitation.

Sonication

Sonication was performed using a Bioruptor Plus (Diagenode, Liège, Belgium) sonicator in cycles of 30 seconds on and 30 seconds off on "high" setting for 30 min. To check that the chromatin was sheared to the desired length (300 bp), a 10 μ l aliquot of sonicated chromatin was reverse cross-linked at 95 °C for 10 min and then 5 μ l was run on 2% agarose e-gel. 30 μ l of 10% Triton X-100 was added to the sonicated lysate and centrifuged at 20,000 rcf for 10 min at 4 °C to pellet debris. The supernatant was transferred to a fresh 2 ml tube and diluted with LB3 containing 1 x

PI to give a total volume of 1 ml with a final concentration of 1% Triton X-100. A 50 μ l aliquot of sonicated whole cell lysate was stored to provide control DNA for input samples.

Chromatin immunoprecipitation

Each mixture containing magnetic bead-bound antibody was added to a separate aliquot of sonicated cell lysate and incubated overnight at 4°C on a rotator.

Elution and cross-link reversal

The magnetic beads were collected using DynaMag-2 (Life Technologies) and the supernatant discarded. In a 4°C cold room, the beads were washed six times with RIPA wash buffer and once with Tris-buffered saline (TBS) (20 mM Tris-HCl, pH 7.6; 150 mM NaCl). Following centrifugation at 960 rcf for 3 min at 4°C, all residual TBS was removed and the beads were suspended in 200 μ l of elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 2% SDS). 150 μ l of elution buffer was added to the thawed 50 μ l input DNA sample.

The ChIPs and input samples were reverse cross-linked by incubation in a water bath at 65°C overnight. The mixture was centrifuged at 16,000 rcf for 1 min at RT and the supernatant was transferred to a new tube.

Digestion of cellular protein and RNA

200 μ l of Tris-EDTA (TE) was added to each tube of ChIP and input DNA. RNA contamination was removed by incubating with 8 μ l RNaseA (1 mg/ml; Ambion, Life Technologies) at 37°C for 30 min. Protein contamination was removed by incubating with 4 μ l proteinase K (20 mg/ml; Invitrogen, Life Technologies) at 55°C for 1 hour.

Phenol-chloroform DNA extraction

Phenol-chloroform extraction was used to further purify the DNA using light Phase Lock Gels (PLGs) (Flowgen Bioscience, Hessle, UK). The PLG was pelleted by centrifugation at 16,000 rcf for 1 min. Equal volumes (400 μ l) of aqueous DNA solution and phenol:chloroform:isoamyl alcohol (P:C:IA, 25:24:1) were added directly to PLG tube, thoroughly mixed, and separated by centrifugation 16,000 rcf for 5 min. The aqueous layer was transferred to a new tube with 2 μ l GlycoBlue (20 μ g/ μ l; Ambion). 16 μ l NaCl was added to a final concentration of 200 mM.

DNA was precipitated by adding 800 μ l of 100% ethanol and incubating for at least 30 min at -80°C. The DNA was then pelleted by centrifugation at 20,000 rcf for 10 min at 4°C, washed with 80% fresh ethanol, dried, and re-suspended in 13 μ l of 10 mM Tris-HCl pH 8.0.

Quantification of ChIP and input DNA

The concentration of DNA in each ChIP and input sample was measured using either a NanoDrop spectrophotometer (Thermo Fisher Scientific) or a Qubit fluorometer (Thermo Fisher Scientific) using the dsDNA high sensitivity assay.

ThruPLEX library preparation

Immunoprecipitated DNA or 50 ng of input DNA was used for library preparation using the ThruPLEX DNA-Seq library preparation protocol (Rubicon Genomics). Briefly, 10 μ l DNA sample was incubated with 3 μ l template preparation mastermix at 22°C for 25 min and then at 55°C for 20 min. Next, 2 μ l library synthesis mastermix was added and incubated at 22°C for 40 min. Libraries were amplified by adding 30 μ l library amplification mastermix and 5 μ l indexing reagent and then incubated as follows: 72°C for 3 min, 85°C for 2 min (extension and cleavage); 98°C for 2 min (denaturation); four cycles of 98°C for 20 s, 67°C for 20 s, 72°C for 40 s (addition of indexes); five to sixteen cycles of 98°C for 20 s, 72°C for 50 s (library amplification, see **Table 2.2**).

Table 2.2 Library amplification cycles

Amount of DNA (ng)	Number of cycles
50	5
20	6
10	7
5	8
2	10
1	11
0.2	14
0.05	16

Library size selection

Adapters were removed and libraries were size selected using AMPure XP beads (Beckman Coulter, Brea, CA) according to the manufacturer's instructions; first, right side selection using (0.5x beads) followed by left-side selection (1.5x beads). 25 μ l AMPure XP beads were added to 50 μ l of library, mixed, and incubated at RT for 10 min, placed on a magnetic stand for 5 min, and the supernatant transferred to a new tube. A further 50 μ l AMPure XP beads were added, incubated for 10 min, placed on a magnetic stand for 5 min, and the supernatant discarded. The beads were washed twice with 200 μ l fresh 80% ethanol and then air dried on the magnet for 15 min. The beads were re-suspended 22.5 μ l 1x TE buffer (pH 8.0), mixed, incubated for 10 min, placed on a magnetic stand, and finally 20 μ l of supernatant was transferred to a fresh tube.

Library quantification

Library fragment size was determined using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA); the ideal library size is 300-400 bp. Library quantification was performed using a QuantStudio 6 Flex Real-Time PCR System (Life Technologies) with the KAPA Library Quantification Kit (Kapa Biosystems; Roche, Basel, Switzerland). 4 μ l sample or DNA standard and 6 μ l KAPA SYBR FAST qPCR Master Mix (ROX low) were added to each well and amplified as follows: 95°C for 20 s followed by 35 cycles of 95°C for 1 s and 60°C for 20 s .

Next-generation sequencing

Pooled libraries were sequenced on an Illumina HiSeq4000 (Illumina, San Diego, CA) according to the manufacturer's instructions to produce single-end 50 bp reads.

2.5.1 Computational analyses of ChIP-sequencing

ChIP-seq data alignment and quality control

Raw sequencing reads from ChIP and input libraries were aligned to the mouse reference genome (GRCm38) using bwa 0.7.12 (Li and Durbin, 2009) backtrack mode with default options. The resulting SAM files were manipulated with samtools 1.3 (Li et al., 2009). Duplicate reads were marked with MarkDuplicates 1.139 from Picard tools (<http://broadinstitute.github.io/picard>).

Quality control (QC) of samples was performed using Phantompeakqualtools (<https://www.encodeproject.org/software/phantompeakqualtools/>; (Marinov et al., 2014)), and only those with a positive quality tag were used in downstream analyses; thus, replicate 1 was removed from the CTCF dataset and replicate 6 from the H3K27ac dataset. All regions within the ENCODE blacklist (<http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm9-mouse/mm9-blacklist.bed.gz>; (Bernstein et al., 2012)) were excluded; the `liftOver` function from the `rtracklayer` 1.34.1 (Lawrence et al., 2009) Bioconductor package was used to convert the coordinates to the mm10 (GRCm38) assembly. Furthermore, any regions with high signal in our own inputs were also excluded; these were identified with the `greyListBS` function from the package `GreyListChIP` 1.6.0 (Brown, 2015) using the merged input datasets.

Differential binding analysis of ChIP-seq data

CTCF differential binding between the two genotypes was performed with `csaw` 1.8.0 (Lun and Smyth, 2016), with a window size of 15 bp, spacing of 50 bp, and a fragment length estimated from cross-correlation analysis. Duplicate reads were retained, but any reads with a mapping quality below 30 were ignored. We checked for evidence of composition biases in the data, but the estimated size factors to correct for this were all very close to one. Thus, count data were normalised for efficiency biases instead. In the differential test we controlled for the batch effect from sample collection time. Windows were merged into regions if they were within 100 bp of each other, restricting the maximum width to 5 kb; this resulted in 42,336 regions. The combined p-value for each region was computed with the Simes' method, upweighting the highest abundance windows (peak summit). Regions with a corrected p-value of 0.05 or lower (FDR <5%) were considered significantly differentially bound. This yielded a set of 787 differentially bound regions, 79.4% of which were less bound in the hemizygous cells. The remaining 162 loci showed a relative enrichment in occupancy compared to the wild-type. Since ChIP-seq quantification is relative in any given sample, the loss of binding in several hundred CTCF binding sites leads to a proportional increase in sequencing reads at other bound loci. We find that these 162 regions are in general of higher affinity (Mann-Whitney test, p-value = 9.13×10^{-24}) and evolutionary conservation (hypergeometric test, p-value = 5.29×10^{-10} ; see below), and longer in width, compared to the stable or less bound binding sites. This is consistent with the compensation expected as a result of the loss of some binding events.

To check the validity of the regions defined as peaks with the csaw method, MACS2 callpeak (Zhang et al., 2008) was run with options `-g mm -s 50 -q 0.01 -call-summits`, using all ChIP libraries merged together along with the corresponding merged inputs. MACS2 reported 47,075 significant peaks, and these contained 97.9% of all csaw regions, verifying that the regions tested for differential binding were significant peaks.

To test for differential binding on the histone data, peaks were called with MACS2 as detailed above. For the H3K4me3 dataset, we kept the option `--call-summits` and for the H3K27ac dataset we instead used `--broad`. Then, DiffBind 2.2.7 (Stark and Brown, 2011) was used to test for differences between the genotypes. Fragment sizes were determined from cross-correlation analysis. To count reads in peaks, we used a summit size of 200 bp for the H3K4me3 dataset and the whole peak for the H3K27ac data. We set `bRemoveDuplicates` to `FALSE` and `mapQcTh` to 30. For differential testing, we controlled for the batch effect from sample collection time and set the options `bSubControl` and `bFullLibrarySize` to `FALSE` when calling `dba.analyze`; we used the edgeR method. For the H3K27ac data, the analysis was performed for two mutually exclusive sets of peaks: those that overlapped an H3K4me3 peak (representing promoters) and those that did not overlap an H3K4me3 peak (putative enhancers).

The MACS2 peak calls are provided as processed data in ArrayExpress, accession code E-MTAB-6261.

Motif analysis on CTCF binding sites

To identify the motifs in the genomic loci occupied by CTCF, the 500 bp DNA sequences centred at the midpoint of the regions defined in the csaw analysis (see above) were extracted. These were then fed to the MEME-ChIP suite (Machanick and Bailey, 2011) for *de novo* motif identification and comparison to the JASPAR Vertebrates and UniPROBE Mouse databases. The most significant motif identified was the canonical CTCF motif (M1), and over 90% of all regions had at least one match. The third most significant motif identified was M2 as defined in Schmidt et al. (2012). For **Figure 3.6B**, one hundred stable and one hundred differential CTCF binding sites were randomly selected. We then collected the coordinate of the M1 motif from the MEME output and extracted the genomic sequences plus 20 nucleotides on both sides. For binding sites with multiple motifs, we selected the one that best matched the motif consensus. The obtained sequences were aligned with MUSCLE (Edgar, 2004) using default parameters.

CTCF motif affinity

To quantify the affinity of each CTCF motif instance identified from our ChIP-seq data, we used DeepBind (Alipanahi et al., 2015), a deep learning algorithm that has been trained on large amounts of ChIP-seq data and that can be used to score the affinity of any given sequence for the CTCF motif. The same 500 bp DNA sequences used for motif discovery (see above) were used to score their motif affinity with DeepBind v0.11 using motif D00328.018 (CTCF). Similar results were obtained if we scored only the motif sequence identified by MEME-ChIP.

Mouse conservation analysis

To investigate whether differentially bound CTCF binding sites have different evolutionary dynamics to stable sites, C57BL/6 CTCF peaks were mapped to their orthologous regions on the genomes of four other mouse species: *Mus musculus castaneus*, *Mus spretus*, *Mus caroli* and *Mus pahari*. This was performed using a multiple whole-genome alignment of 17 eutherian mammals (Herrero et al., 2016) plus *mcast*, *mspr*, *mcar*, and *mpah* (Thybert et al., 2018). A CTCF peak was defined as conserved across all five mouse species if its orthologous locus in each species was also proven to bind CTCF based on ChIP-seq data derived from that species (Thybert et al., 2018). Significant depletion of conserved peaks in the set of differentially bound CTCF sites was tested for using a hypergeometric test.

2.6 RNA-sequencing

Cell lysis and RNA isolation

Total RNA was extracted from P4 MEF cultures from six biological replicates from each genotype using QIAzol Lysis Reagent (Qiagen) according to the manufacturer's instructions. 700 μ l QIAzol Lysis Reagent was added to a dish containing ~500,000 cells, incubated at RT for 5 minutes. 140 μ l chloroform was added, and the sample was mixed vigorously for 15 sec, incubated at RT for 2 min, and then transferred to a PLG heavy tube (Flowgen Bioscience). Aqueous and organic phases were separated by centrifugation at 12,000 rcf for 10 min; the aqueous phase was transferred to a new tube, mixed with a 1 x volume of 100% isopropanol, and incubated at RT for 15 min. RNA was precipitated by centrifugation at 20,000 rcf for 20 min at 4°C. The RNA

pellet was washed twice with fresh 60% ethanol, dried, and resuspended in RNase-free water. RNA concentration was measured using a Nanodrop spectrophotometer (Thermo Fisher). RNA quality was assessed on a Total RNA Nano Chip Bioanalyzer (Agilent).

DNase treatment and removal

DNase treatment and removal was performed using the TURBO DNA-free™ Kit (Ambion, Life Technologies) according to the manufacturer's instructions. 10 µg of total RNA was incubated with 0.1 volume 10X DNase Buffer (Life Technologies) and 1 µl TURBO DNase (Life Technologies), mixed gently, and incubated at 37°C for 30 min. 0.1 volume DNase Inactivation Reagent (Life Technologies) was added, mixed, and incubated at RT for 5 min. The sample was centrifuged at 10,000 rcf for 90 sec and the supernatant transferred to a new tube. RNA was precipitated by adding 6 volumes of 100% ethanol and sodium acetate to a final concentration of 75 mM and incubating at -20°C for >1 h. RNA was pelleted by centrifugation at 10,000 rcf for 30 min at 4°C and washed twice with fresh 70% ethanol.

Library preparation

RNA-sequencing libraries were generated using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina) according to the manufacturer's instructions.

rRNA depletion: 1 µg of RNA was diluted with nuclease-free ultra pure water to a final volume of 10 µl. 5 µl of rRNA Binding Buffer (Illumina) and 5 µl of Ribo-Zero Gold rRNA Removal Mix (Illumina) were added and incubated at 68°C for 5 min and then at RT for 1 min. The sample was transferred to new tube containing 35 µl rRNA Removal Beads (Illumina), mixed thoroughly, placed on a magnetic stand at RT for 1 min, and then the supernatant was transferred to a new tube. 99 µl RNAClean XP beads (Illumina) beads were added, mixed, incubated at RT for 15 min and then for a further 5 min on a magnetic stand. The supernatant was discarded, the beads were washed with fresh 70% ethanol, air dried, and then RNA was eluted in 11 µl Elution Buffer (Illumina). The sample was returned to the magnet, and 8.5 µl supernatant was transferred to a new tube containing 8.5 µl Elute, Prime, Fragment High Mix (Illumina) and incubated at 94°C for 8 minutes.

First strand cDNA synthesis: 0.8 μ l SuperScript II and 7.2 μ l First Strand Synthesis Act D Mix were added to the RNA sample. The synthesis reaction was performed as follows: 25°C for 10 min, 42°C for 15 min, 70°C for 15 min.

Second strand cDNA synthesis: 5 μ l of Resuspension Buffer and 20 μ l Second Strand Marking Master Mix was added to each sample and incubated at 16°C for 1 h. 90 μ l AMPure XP beads were added to each sample, mixed, incubated at RT for 15 min, and then placed on a magnetic stand for 5 min. The supernatant was discarded, beads were washed twice with 200 μ l fresh 80% ethanol, dried, resuspended in 17.5 μ l Resuspension Buffer, incubated at RT for 2 min, and then returned to the magnetic stand. 15 μ l supernatant (containing double-stranded cDNA) was then transferred to a new tube.

Adenylate 3' ends: 2.5 μ l of Resuspension Buffer and 12.5 μ l A-Tailing Mix was added to each sample and incubated at 37°C for 30 min and then 70°C for 5 min.

Ligate adapters: 2.5 μ l of Resuspension Buffer and 2.5 μ l of Ligation Mix and 2.5 μ l unique RNA adapters were added to each sample and incubated at 30°C for 10 min. 5 μ l Stop Ligation Buffer was added. Clean up was performed by adding 42 μ l AMPure XP beads to each sample, incubating at RT for 15 min, placing on a magnetic stand for 5 min, and the supernatant was then discarded. The beads were washed twice with fresh 80% ethanol, air dried, and resuspended in 52.5 μ l Resuspension Buffer. After returning to the magnetic stand, 50 μ l supernatant was transferred to a new tube containing a further 50 μ l AMPure XP beads for a second clean up. The final resuspension volume was 22.5 μ l, and 20 μ l of cleaned up sample was transferred to a new tube.

Enrich DNA fragments: 5 μ l PCR Primer Cocktail and 25 μ l PCR Master Mix were added to each sample and incubated as follows: 98°C for 30 sec; 15 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 30 s; and 72°C for 5 min. Libraries were then cleaned up using AMPure XP beads, as above, with the initial bead volume of 47.5 μ l, resuspension in 32.5 μ l Resuspension Buffer, and 30 μ l of clean, enriched library was transferred to a new tube.

Library quantification

Library fragment size was determined using a 2100 Bioanalyzer (Agilent). Libraries were quantified by qPCR (Kapa Biosystems) and then pooled to ensure an equal balance of libraries in the final pool.

Next generation sequencing

Pooled libraries were sequenced on a HiSeq4000 according to manufacturer's instructions to produce paired-end 150 bp reads.

2.6.1 Computational analyses of RNA-sequencing

RNA-seq data processing and analysis of (MEF libraries)

RNA-seq paired-end fragments were aligned to the mouse reference genome (GRCm38) with STAR 2.5.2a (Dobin et al., 2013) with options:

```
--outFilterMismatchNmax 6 --outFilterMatchNminOverLread 0.4  
--outFilterScoreMinOverLread 0.4 --outFilterType BySJout  
--outFilterMultimapNmax 20 --alignSJoverhangMin 8  
--alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 1000000  
--alignMatesGapMax 1000000 --outSAMstrandField intronMotif.
```

On average, 77% of the sequencing fragments mapped uniquely. The numbers of fragments overlapping annotated transcripts were obtained with featureCounts 1.5.2 (Liao et al., 2014) from the Subread package using Ensembl's genome annotation (Aken et al., 2017) version 84 (http://mar2016.archive.ensembl.org/Mus_musculus/Info/Index). The sequence of the *lacZ* cassette was added to the genome to corroborate the genotype of each sample.

For differential expression analysis, hidden batch effects were identified with the Bioconductor package sva 3.22.0 (Leek et al., 2017), providing known batch effects (sample collection time). We then used DESeq2 1.14.1 (Love et al., 2014) to test for differential expression, controlling for both the known and hidden batch effects. Genes were considered significantly differentially expressed if their adjusted p-value was lower than 0.05 (equivalent to a false discovery rate of 5% or less).

Significantly enriched gene ontology terms were identified using the functions goana and kegg from the Bioconductor package limma 3.33.14 (Ritchie et al., 2015), with gene length as a covariate. The gene lengths supplied were obtained from the featureCounts output (see above). The reported p-values were corrected for multiple

testing by the Benjamini-Hochberg method and were considered significant if they were lower than 0.05. Detailed results of both the differential expression analyses are provided in **Appendix A**. The raw and normalised RNA-seq counts are provided as processed data in ArrayExpress, accession code E-MTAB-6259.

RNA-seq data processing and analysis (tumour libraries)

RNA-seq paired-end fragments were aligned to the mouse reference genome (GRCm38) with STAR 2.5.2a (Dobin et al., 2013). Transcripts abundances were quantified using kallisto v0.43.1 (Bray et al., 2016), which uses pseudoalignment to rapidly determine the compatibility of reads with targets without the need for alignment. RNA-seq libraries from DEN-initiated tumours were grouped according to genotype. All 0 values were replaced with 0.01 (giving a minimum \log_2 value of -6.6). The \log_2 fold-change was calculated by the median of the \log_2 values of the *Ctcf*^{+/-} group minus the median of the \log_2 values of the wild-type group. A Mann–Whitney U test was performed using the \log_2 values of each group; transcripts with equal distribution were removed from further analyses. Multiple correction testing (q values) were calculated using the Benjamini-Hochberg method. **Figure 4.16** was generated using $-\log_{10}(\text{q value})$.

RNA-seq data processing and analysis (published mouse tumour libraries)

The RNA-seq data from mouse liver and tumour samples (Connor et al., 2018) was processed as detailed above but using the C3H/HeJ genome as a reference (ftp://ftp.ensembl.org/pub/release-89/fasta/mus_musculus_c3hhej/dna/Mus_musculus_c3hhej.C3H_HeJ_v1.dna_sm.toplevel.fa.gz) (Lilue et al., 2018). To test for differential expression between the normal liver and tumour samples we used DESeq2 (Love et al., 2015) and genes were considered significantly differentially expressed if their adjusted p-value was lower than 0.05 (FDR < 5%).

To compare to the list of differentially expressed genes in the *Ctcf* hemizygous MEFs, we matched genes by their official gene name. Genes that were significantly differentially expressed in both datasets were deemed concordant if they were up- or down-regulated both in the *Ctcf* hemizygous MEFs and in the tumours, compared to their respective controls. Gene set enrichment analysis was performed as detailed previously, using only the set of concordant differentially expressed genes as the test set. The raw and normalised RNA-seq counts of the mouse liver tumour and

control datasets are provided as processed data in ArrayExpress, accession codes E-MTAB-6971 and E-MTAB-6972.

RNA-seq data processing and analysis (published human tumour libraries)

To compare the set of dysregulated genes in the *Ctcf* hemizygous MEFs to alterations in the transcriptomes of human cancers, we mined The Cancer Genome Atlas PanCanAtlas to obtain a list of uterine and breast tumour samples with identified missense, frameshift or stop-gain mutations in *CTCF*. We collected the RNA-seq raw counts for these samples, along with all available control normal uterine and breast tissue samples. The gene annotation used was from <https://www.gencodegenes.org/releases/22.html>. We used DESeq2 (Love et al., 2015) to normalise and test for differential expression between the tumour and control samples, for each tissue separately.

To compare these results to the genes altered in the *Ctcf* hemizygous MEFs, we obtained the orthology relationships between the human and mouse genome using Ensembl version 84 (Herrero et al., 2016) and restricted our analysis to one-to-one orthologs. Genes that were significantly differentially expressed in both datasets were deemed concordant if they were up- or down-regulated both in the *Ctcf* hemizygous MEFs and in the tumours, compared to controls. Gene set enrichment analysis was performed as detailed previously, using only the set of concordant differentially expressed genes as the test set.

2.7 TMT proteomics

Tissue homogenisation

MEF cultures for five biological replicates of each genotype were washed with ice cold PBS. Cells were lysed in 200 μ l of 0.1 M TEAB, 0.1% sodium dodecyl sulphate (SDS) at 90°C for 10 min, followed by tip sonication. Total protein was quantified using a Bradford assay (Bio-Rad Quick Start; Bio-Rad Laboratories, Hercules, CA) according to the manufacturer's instructions.

Preparing whole cell protein extracts

90 μg of protein per sample was reduced by the addition of 2 μl 50 mM tris-2-carboxyethyl phosphine (TCEP) for 60 min at 60°C followed by cysteine blocking for 10 min at RT using 1 μl 200 mM methyl methanethiosulfonate (MMTS).

Protein digestion

5 μl of 375 mM iodoacetamide was added to the sample and incubated for 30 min in the dark. Six volumes of pre-chilled acetone were added for overnight precipitation. The sample was centrifuged at 8000 rcf for 10 min at 4°C, the supernatant discarded, and the pellet air-dried for 10 min. Trypsin digestion (protein/trypsin ratio 30:1) was performed overnight at 37°C.

Peptide labelling

0.8 mg tandem mass tag (TMT) 10-plex label reagents (Thermo Fisher Scientific) were added to 41 μl of ethanol and incubated at RT for 5 min with intermittent vortexing. This was added to the peptide solution and incubated for 1 h at RT. 8 μl of 5% hydroxylamine was added to the sample and incubated for 15 min to quench the reaction. Equal amounts of the ten TMT-labelled samples were pooled for multiplex analysis.

Liquid chromatography-mass spectrometry analysis

The TMT mixture was then basic reverse phase (bRP) fractionated on a Dionex Ultimate 3000 system at high pH using the X-Bridge C18 column (3.5 μm 2.1 x 150 mm, Waters Corp., Milford, MA).

Fractions were analysed on a Dionex Ultimate 3000 UHPLC system coupled with the nano-ESI Fusion Lumos (Thermo Fisher Scientific). Samples were loaded on the Acclaim PepMap 100, 100 μm x 2cm C18, 5 μm , 100Å trapping column with the ulPickUp injection method using the loading pump at 5 $\mu\text{l}/\text{min}$ flow rate for 10 min. For peptide separation, the EASY-Spray analytical column 75 μm x 25 cm, C18, 2 μm , 100Å was used for multi-step gradient elution. Mobile phase was composed of 2% acetonitrile, 0.1% formic acid and mobile phase was composed of 80% acetonitrile, 0.1% formic acid. The Lumos was operated in data-dependent mode for both MS2 and Synchronous Precursor Selection (SPS)-MS3 methods. The full scan was performed in the Orbitrap in the range of 380-1500 m/z at 120 K

resolution and the MS2 scan was performed in the ion trap with collision energy 35%. Peptides were isolated in the quadrupole with isolation window 0.7 Th. The 10 most intense fragments were selected for SPS HCD-MS3 analysis with MS2 isolation window 2.0 Th. The HCD collision energy was set at 55%, and the detection was performed with Orbitrap resolution 60 K and in scan range 100-400.

2.7.1 Computational analyses of proteomic data

Raw data were processed with the Sequest HT search engine in Proteome Discoverer 2.1 software. All spectra were searched against a UniProtKB/Swiss-Prot fasta file containing 16,915 reviewed mouse entries. The parameters for the SequestHT node were as follows: precursor mass tolerance 20 ppm; fragment mass tolerance 0.5 Da; dynamic modifications were oxidation of M (+15.995 Da), deamidation of N, Q (+0.984 Da); and static modifications were TMT6plex at any N-terminus, K (+229.163 Da) and methylthio at C (+45.988). The consensus work flow included S/N calculation for TMT intensities as previously described (McAlister et al., 2014), and the level of confidence for peptide identifications was estimated using the Percolator node.

Peptide intensity data were quantile normalised and summarised into protein-level counts by summing the intensity values for all peptides for a given protein. Samples were inspected via hierarchical clustering and principal component analysis (PCA) to identify outliers of low quality; these were removed from downstream analyses (WT replicates 2 and 3 and *Ctcf* hemizygous replicates 1 and 3). Limma 3.33.14 (Ritchie et al., 2015) was used to assess differential protein expression between the genotypes, controlling for batch effects (sample collection time).

2.8 Hi-C

Hi-C was performed according to published protocols (Harewood et al., 2017; Lieberman-Aiden et al., 2009; Nagano et al., 2015).

MEF cultures (~20 million cells) for three biological replicates of each genotype were fixed in DMEM containing 2% fresh formaldehyde and incubated at RT for 10 min, quenched with 1 M glycine for 5 min, and washed twice with ice cold PBS. The fixed cells were lifted off the plate, pelleted by centrifugation (400 rcf at 4°C for 10 min), and flash-frozen at -80°C.

Lysis: 1 ml ice-cold lysis buffer (10 mM Tris-HCl, 0.2% NP-40/Igepal CA-420, 10 mM NaCl, and 1x protease mini inhibitor cocktail tablet (EDTA-free)) were added to the cell pellet and transferred to a 2 ml glass homogeniser. The sample was dounced up and down 10 times, incubated on ice for 1 min, dounced a further 10 times, and then incubated on ice for 30 min with occasional mixing. The nuclei were pelleted by centrifugation, the supernatant discarded, and nuclei resuspended in 358 μ l ice cold 1.25x NEBuffer 2 (New England BioLabs, Ipswich, MA). To remove proteins that were not directly cross-linked to DNA, 11 μ l 10% SDS was added and incubated at 37°C for 60 min, shaking at 950 rpm. The reaction was quenched by adding 75 μ l 10% Triton X-100 and incubated at 37°C for 60 min, shaking at 950 rpm.

HindIII digest: Chromatin was digested by adding 1000 units of HindIII (10 μ l of 100 u/ μ l NEB R0104T) and incubating at 37°C overnight while rotating (950 rpm).

Biotinylation of DNA ends: To fill in the restriction fragment overhangs and mark the DNA ends with biotin, the following was added to each tube: 1.5 μ l 10 mM dCTP, 1.5 μ l 10 mM dGTP, 1.5 μ l 10 mM dTTP, 37.5 μ l 0.4 mM biotin-14-dATP (Invitrogen 19524-016), 10 μ l 5 u/ μ l Klenow (DNA polymerase I large fragment, NEB M0210L). This was then mixed and incubated at 37°C for 60 min, shaking at 700 rpm for 10 sec in every 30 sec.

Ligation: 100 μ l T4 ligase reaction buffer (NEB B0202S), 10 μ l BSA (10 mg/ml, NEB B9001S), and 354 μ l dH₂O was added to each sample and 1 u/ μ l T4 DNA ligase (Invitrogen 15224-025) was added to each Hi-C tube and incubated at 16°C for 4 h and 12°C overnight.

Excess lysis buffer removal: Nuclei were pelleted by centrifugation and 500 μ l of lysis buffer was removed and discarded.

Proteinase K: Cross-linking was reversed and protein was degraded by adding 10 μ l 10 mg/ml Proteinase K (Roche 03115879001) per tube and incubating the tubes all day and overnight at 65°C.

DNA purification: 1.8x volume AMPure XP beads were added to the 500 μ l Hi-C sample, mixed, incubated at RT for 15 min, washed twice with fresh 80% ethanol,

dried, DNA eluted in 30 μ l TLE, and transferred to a new tube.

Quantification: DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific).

Hi-C ligation efficiency control: To detect known short- and long-range interactions in the 3C and Hi-C libraries, PCR was performed using Hotstart Polymerase Master Mix (5 μ l 5x PCR mix, 1 μ l of each 10 μ M primer, 0.5 μ l/2.5u Qiagen HotStar Taq DNA polymerase) and 100 ng template in a 25 μ l reaction volume. The following PCR protocol was used: 95°C for 15 min; 35 cycles of 94°C for 30 sec, 53°C for 30 sec, 72°C for 1 min; 72°C for 10 min; 12°C forever. The PCR product was run on a 0.8% agarose gel; both libraries were expected to run as a rather tight band over 10 kb. The primers for the short-range interaction were expected to span more than one restriction fragment and/or be in the same orientation.

Hi-C biotin marking and ligation efficiency was verified by a PCR digest assay. Successful fill-in and ligation of a HindIII site (AAGCTT) created a site for the restriction enzyme NheI (GCTAGC). 25 μ l PCR reactions were set up to amplify a short-range ligation product formed from two nearby restriction fragments (AlbF1 and AlbF2). PCR products were purified using Qiagen Minelute PCR purification kit according to the manufacturer's instructions and eluted in 50 μ l water. The purified sample was then divided into four 12 μ l aliquots: undigested, digested with HindIII, digested with NheI, and digested with both HindIII and NheI, and then incubated at 37°C for 2 h. The HindIII/NheI digestion was: 12 μ l DNA, 2 μ l Cutsmart buffer, 3 μ l water, 3 μ l HindIII-HF or NheI-HF (20,000 u/ml). The combined HindIII/NheI digestion was: 12 μ l DNA, 2 μ l Cutsmart buffer, 1 μ l water, 2.5 μ l HindIII-HF (20,000 u/ml), 2.5 μ l NheI-HF (20,000 u/ml). Samples were run on a 1.5% agarose gel.

Removal of biotin at non-ligated DNA ends: To avoid pulling down non-ligated biotinylated fragments, biotin-dATP was removed from these un-ligated ends using the exonuclease activity of T4 DNA polymerase.

5 μ g Hi-C library was mixed with 1 μ g 10 mg/ml BSA, 10 μ l 10x NEB buffer 2, 2 μ l 10 mM dATP, and 15 u (5 μ l) T4 DNA polymerase (NEB M0203L). The volume was adjusted to 100 μ l with water, and the mixture was incubated at 20°C for 2 h. The reaction was stopped by adding 2 μ l 0.5M EDTA pH 8.0. DNA was purified using 1.8x AMPure XP beads as described above.

DNA shearing: A Covaris S220 (Covaris, Woburn, MA) was used to shear the DNA: duty factor 5%, peak incident power 175 W, cycles per burst 200, for 55 seconds.

End Repair: The following was added to each sample: 25 μ l 10x NEB2, 2.5 μ l 10 mM dATP, 12.5 μ l 2 mM dNTP mix, 30 μ l water, 10 μ l T4 DNA polymerase, 10 μ l T4 PNK, 10 μ l Taq DNA polymerase, and then incubated at 25 °C for 20 min and at 72 °C for 20 min.

Double-sided SPRI selection To select fragment sizes 200-700 bp (majority 250-550 bp), sequential solid phase reversible immobilisation (SPRI) selection was performed, first by adding 0.6x volume AMPure XP beads to the 230 μ l DNA sample, incubating as described above, and then transferring the supernatant to a new tube. A second size selection step was performed using 1x volume beads, incubating, washing the beads twice in fresh 80% ethanol, drying, and eluting DNA in 30 μ l TLE. The DNA was quantified using a Qubit fluorometer.

Biotin pull-down: Using LoBind tubes (Eppendorf) and tips (STARLAB), 25 μ l Dynabeads MyOne Streptavidin C1 beads (Invitrogen) were washed twice with 400 μ l Tween Buffer (TB, 5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween) and resuspended in 30 μ l 2x No Tween Buffer (2x NTB, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 2 M NaCl). The 30 μ l Hi-C DNA was added, mixed, and incubated at RT for 15 min. Beads were reclaimed, the supernatant discarded, and beads washed in 400 μ l 1x No Tween Buffer (1x NTB, 5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl). Beads were washed with 100 μ l dH₂O, resuspended in 19 μ l water, and transferred to new tube.

Adapter ligation: Adapters were ligated to the Hi-C library by the addition of 1 μ l 15 μ M stock adapter, 20 μ l 2x Quick Ligation Buffer (NEB M2200), and 2 μ l Quick T4 DNA Ligase (2,000,000 units/ml, NEB M2200), and incubating at RT for 15 min. The beads were recaptured, supernatant discarded, beads washed twice with 400 μ l TB, once with 200 μ l 1x NTB, once with 100 μ l, and then 50 μ l 1x NEB2, before finally being resuspended in 25 μ l 1x NEB2 and transferred to a new tube.

PCR amplification: Each 25 μ l amplification reaction contained: 2.5 μ l Hi-C beads, 1.5 μ l 10 μ M TruSeq PCR primer mix, 2.5 μ l 2mM dNTP mix, 5 μ l 5x HF Buffer, 0.3

μ l NEB Phusion, and 13.2 μ l water. The mixture was then amplified as follows: 98°C for 30 s; seven cycles of 98°C for 10 s, 65°C for 30 s, 72°C for 30 s; 72°C for 7 min.

The PCR products were purified using AMPure XP beads as described above using 0.7x bead volume and elution in 50 μ l TLE.

Library quantification

Library fragment size measurement and quantification were performed using a 2100 Bioanalyzer (Agilent). Libraries were pooled to ensure equal representation in the final sequencing pool.

Next-generation sequencing

Pooled libraries were sequenced on an Illumina HiSeq4000 according to the manufacturer's instructions to produce paired-end 150 bp reads.

2.8.1 Computational analyses of Hi-C data

Each sample was sequenced to a mean depth of ~179 million paired-end reads totalling over a billion read pairs for the complete dataset. Data were mapped and QCed with HiCUP 0.5.8 (Wingett et al., 2015) and bowtie2-2.2.8 (Langmead and Salzberg, 2012) using the GRCm38 mouse reference genome. Over 60% of all read pairs were properly mapped and paired; from these, over 85% were valid pairs and the uniqueness percentage after de-duplication was ~70%. The BAM files produced by HiCUP, which contain only valid, non-redundant read pairs, were used for downstream analyses.

The HiCUP output BAM files were converted to a format compatible with HOMER using the hicup2homer utility. Using HOMER (Heinz et al., 2010), tag directories were created from the merged data of the three wild-type or *Ctcf* hemizygous samples. The correlation between the interaction profiles of the two genotypes was calculated with the getHiCcorrDiff.pl script using a resolution of 100 kb and a super-resolution of 150 kb.

To identify chromatin loop interactions, the analyzeHiC program from HOMER (Heinz et al., 2010) was used on the merged data from all six replicates; this ensured that the definition of significant interactions was agnostic to the genotype, allowing us to subsequently perform differential analysis of the wild-type and hemizygous profiles

without compromising FDR control (Lun and Smyth, 2014). We supplied analyze-HiC with the options `-res 20000 -interactions -nomatrix -maxDist 10000000 -minDist 5000 -center`. Differential analysis on the identified loops was performed using diffHic (Lun and Smyth, 2015) on the set of interactions reported by HOMER with an FDR lower than 0.05 and restricted to the autosomes. The HiCUP output BAM files were processed with the `preparePairs` function, keeping data for each replicate separate; any fragments mapping against the blacklisted regions used for the ChIP-seq analyses (see above) were discarded. Then, the function `connectCounts` was used to count the number of fragments mapping specifically to the loci involved in the loop interactions. Only interactions that had more than 20 average counts per million (90,704 loops) were used for differential testing. Data were normalised for depth of sequencing by providing the library sizes of the complete dataset. Differential testing was performed, controlling for the batch effect from sample collection time, and the resulting p-values were corrected for multiple testing by the Benjamini-Hoechberg procedure. Finally, we used this ranked list to test whether looping interactions overlapping with differentially expressed genes (plus 5 kb on either side) or differentially bound CTCF sites were enriched at the top of the list using the function `geneSetTest` (Wilcoxon signed rank test) from the `limma` package.

Definition of gene-enhancer pairs

To determine if the enhancers likely to regulate differentially expressed genes changed with gene expression, we retrieved all putative enhancer peaks (defined as H3K27ac peaks that did not overlap with H3K4me3 peaks) that were linked to a differentially expressed gene via a significant interaction in the Hi-C data (see above). For each of the 296 dysregulated genes, 261 had at least one and up to 65 linked enhancers (median = 8). Only a subset of these gene-enhancer pairs were likely to be *bona fide* regulatory interactions. To increase our signal-to-noise ratio, we reasoned that we could use the paired nature of our datasets to infer correlations between the RNA expression levels and H3K27ac abundance, since both measurements were performed in the same MEF cultures. Thus, for each gene-enhancer pair, we calculated the Pearson correlation coefficient between the RNA-seq- and ChIP-seq-normalised counts for the five replicates that had successful libraries for both methodologies. We then retained the gene-enhancer pair with the highest correlation value for each dysregulated gene. **Figure 3.8** was plotted with these

pairings; for genes with no linked enhancer, the corresponding row in the heatmap has been left blank.

To generate the heatmaps shown in **Figure 3.8**, we defined 1 kb windows centred either at the transcription start site (as defined in Ensembl v84) or the midpoint of the H3K27ac peak, extending 17 kb up- and downstream. The number of sequencing reads mapping to such windows in the histone ChIP-seq data were obtained with bedtools v2.24.0, command `intersect -c` (Quinlan and Hall, 2010). The counts for each sample were normalised to account for the total depth of sequencing and then aggregated into 5 kb bins. For each 5 kb bin, the average abundance across all replicates was used, and the \log_2 fold-change between the genotypes was plotted.

2.9 Mouse tumour models

2.9.1 Spontaneous tumourigenesis

A cohort of *Ctcf*^{+/-} and wild-type mice were aged to 20 months. Full necropsy was performed and all tissues processed for histological examination. In addition, all macroscopically identified tumours >2 mm were bisected and processed for parallel DNA/RNA extraction and histopathology (see below).

2.9.2 Chemical model of hepatocarcinogenesis

15-day-old male offspring from *Ctcf*^{+/-} x *Alb-cre* breeding pairs were treated with a single intraperitoneal injection of DEN (Sigma-Aldrich N0258; 20 mg/kg body weight) diluted in 0.85% saline. Ear biopsies were taken immediately prior to treatment, which was therefore blinded to the genotype (hepatocyte-specific *Ctcf* hemizygous or wild-type).

In a pilot experiment, to determine the optimal time point for sample collection, two mice of each genotype were euthanised every two to four weeks from 19 to 35 weeks after treatment with DEN.

Subsequently, liver tumour samples were collected from DEN-treated mice 36 or 42 weeks after treatment. Mouse body and organ weights were recorded. Full necropsies were performed, livers photographed, all macroscopically identified tumours measured, and those >2 mm in diameter bisected and processed in parallel for DNA/RNA extraction and histopathological examination (see below) along with background "normal" tissues.

2.10 Tissue collection and processing

2.10.1 Fresh frozen tissue

Ear and tail samples from all mice were flash frozen at the time of necropsy and used for genotyping and/or as normal genome control samples. Liver tumours were macroscopically identified and isolated. Nodules of sufficient size (>2 mm diameter) were bisected; one half was flash frozen in liquid nitrogen and stored at -80°C for RNA/DNA extraction, and the other half was processed for histology (see below). Background non-tumour liver was sampled from all mice, again flash frozen and fixed for histology.

DNA and RNA isolation

Simultaneous isolation of genomic DNA (gDNA) and total RNA from liver tissue and liver tumours was performed using the AllPrep 96 DNA/RNA Kit (Qiagen, 80311) according to the manufacturer's instructions. Briefly, 10 mg tissue was homogenised in 350 µl Buffer RLT using a TissueLyser II (Qiagen) at 25 Hz for 2 min, centrifuged at 5600 rcf for 4 min at RT, and the supernatant transferred to the wells of the AllPrep 96 DNA plate. This plate, containing spin columns, was then centrifuged again at 5600 rcf for 4 min at RT. The flow-through from this plate contained total RNA for purification, as detailed below. The genomic DNA, still on the membranes in the spin column plate, was purified as described.

Genomic DNA purification: 800 µl Buffer AW1 was added to each well of the plate, centrifuged at 5600 rcf for 4 min at RT, then 800 µl Buffer AW2 was added and the centrifugation step was repeated for 10 min to dry the membranes. The DNA was eluted by adding 70 µl of pre-warmed Buffer EB, incubated for 5 min, and centrifuged at 5600 rcf for 4 min at RT. The elution step was repeated with a further 70 µl EB to ensure complete recovery of DNA.

QC was performed by running 1 µl of gDNA (diluted 1:20) on a 1% agarose gel. DNA quantification was performed in triplicate using the Quant-IT dsDNA Broad Range Kit (Thermo Fisher Scientific).

Total RNA purification: 350 µl 70% ethanol was added to each well containing the flow-through, transferred to an RNeasy 96 plate, and centrifuged at 5600 rcf for 4 min at RT. 800 µl Buffer RW1 was added to each well of the plate containing

spin columns, centrifuged at 5600 rcf for 4 min at RT, then 800 μ l Buffer RPE was added to each well, centrifuged again, and finally another 800 μ l Buffer RPE was added to each well and centrifuged for 10 min to dry the membranes. RNA was eluted by adding 50 μ l RNase-free water to each well then incubated for 1 min at RT and centrifuged at 5600 rcf for 4 min at RT. A further 50 μ l RNase-free water was added, incubated, and the plate was centrifuged again to ensure complete recovery of RNA.

2.10.2 Fixed tissue for histology

Tissue samples were fixed in 10% neutral buffered formalin for 24 h, transferred to 70% ethanol, machine processed (Leica ASP300 Tissue Processor; Leica, Wetzlar, Germany), and paraffin embedded. All formalin-fixed paraffin-embedded (FFPE) sections were 3 μ m in thickness.

Histochemical staining

FFPE tissue sections were stained with haematoxylin and eosin (H&E) for morphological assessment and using Gomori's method for reticular fibres to assess liver architecture. Histochemical staining was performed using the automated Leica ST5020; mounting was performed on the Leica CV5030. Briefly, for H&E staining: sections were de-waxed with xylene (2 x 10 min), rehydrated with absolute ethanol (2 x 5 min), 70% ethanol (1 x 5 min), briefly washed in water, and stained with Harris haematoxylin (5 min) before washing in running tap water (5 min). Differentiation was performed with 1% acid alcohol (20 s) and then washing in tap water (5 min). Sections were counterstained with 1% aqueous eosin (5 min), washed in tap water (10 s), dehydrated through 50% ethanol (20 s), 70% ethanol (20 s), 100% ethanol (30 s), 100% ethanol (1 min), cleared with xylene (2 x 5 min), and mounted in DPX.

Immunohistochemistry

Immunohistochemistry was performed on FFPE tissue sections with antibodies targeting Ki67, CD31, CD45, β -catenin, CK8, and CK14 (**Table 2.3**) using the Bond Polymer Refine Detection Kit (DS9800, Leica Microsystems) with DAB enhancer (Leica Biosystems, AR9432) on the automated Bond platform. Heat-induced epitope retrieval (HIER) using sodium citrate or Tris-EDTA pre-treatments were run at 100 °C on the Bond platform. Proteolytic-induced epitope retrieval (PIER) enzyme digestion

was run at 37°C on the Bond platform using Leica's Bond enzyme concentrate (AR9551) containing a proteolytic enzyme (17 mg/ml) and stabiliser (6 µl/ml). A mouse-on-mouse protocol was used for β -catenin using the Polymer Refine template with an additional IgG block (Vector, MKB-2213) and an isotype-specific secondary antibody (rabbit anti-mouse IgG, Abcam, ab125913, diluted 1:1500). For the antibodies raised in rat, the post primary was substituted for a rabbit anti-rat antibody (Bethyl Laboratories, Montgomery, TX; reference A110-322A, diluted 1:250).

Table 2.3 Antibody details and conditions used for immunohistochemistry.

Target	Manufacturer & cat. no.	Isotype	Clonality	Dilution	Retrieval
Ki67	Bethyl Laboratories, 00375	Rabbit IgG	Polyclonal	1:1000	Sodium citrate, 20'
CD31	BD Biosciences, 553370	Rat IgG	Monoclonal	1:50	Enzyme digestion, 10'
CD45	R&D Systems, MAB1217	Rat IgG	Monoclonal	1:750	Sodium citrate, 10'
β -catenin	BD Biosciences, 610154	Mouse IgG	Monoclonal	1:100	Sodium citrate, 20'
CK8	DSHB, TROMA-1	Mouse IgG	Monoclonal	1:100	Sodium citrate, 20'
CK14	BioLegend, 905301	Rabbit IgG	Polyclonal	1:5000	Sodium citrate, 10'

Imaging

All tissue sections were scanned using the Aperio XT system (Leica Biosystems) at 20x resolution. H&E-stained tissue section images of the tumours included in whole genome sequencing experiments are available at BioStudies archive at EMBL-EBI under accession S-BSST129.

2.11 Tumour histopathology

H&E sections of liver tumours were reviewed (n = 1648). Tumour size, morphological subtype, presence of portal tracts, mitotic index, and the presence of cystic change, haemorrhage, necrosis, or vascular invasion were assessed.

In addition, histochemical stains and IHC were performed on a subset of tumours to aid tumour classification and also on all tumours selected for whole genome sequencing.

Tissue sections were blinded and assessed twice; discordant results were reviewed by an independent clinical hepatobiliary pathologist (Dr Susan Davies). Tumours were classified according to the International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice (INHAND) guidelines

(Thoolen et al., 2010). Dysplastic nodules have an expansile growth pattern causing compression of adjacent hepatic parenchyma, loss of normal lobular architecture (irregular reticulin fibre staining), nuclear atypia, and may show increased proliferation (increased Ki67 staining). Hepatocellular carcinomas are characterised by thickened trabeculae (loss of reticulin fibre staining), pseudoglandular structures, more marked cellular atypia, increased nuclear to cytoplasmic ratios, higher proliferative index (markedly increased Ki67 staining), and an infiltrative growth pattern.

Sample selection for WGS

A total of 96 samples were selected for whole genome sequencing. Whenever possible, a paired normal liver sample and tail/ear from the same mouse was included in addition to the tumours. All spontaneous tumours were selected. Since carcinogen-induced tumours arising in the same liver are independent (Connor et al., 2018), multiple tumours were selected from each mouse to optimise the number of tumour samples sequenced. Six mice were randomly selected from those with 4 to 14 tumours (mean \pm SD) that included both DN and HCCs. Of these, four to six tumours per mouse were selected that met the following histological criteria: (1) diagnosis of either DN or HCC, (2) homogenous tumour morphology, (3) tumour cell percentage $>80\%$, and (4) adequate tissue for DNA extraction. Neoplasms with extensive necrosis, mixed tumour types, a nodule-in-nodule appearance (indicative of an HCC arising within a DN), or contamination by normal liver tissue were excluded.

2.12 Whole genome sequencing

Library preparation

1 μ g of 50 ng/ μ l high molecular gDNA was used to generate sequencing libraries using the TruSeq PCR-free Library Prep Kit (Illumina) according to the manufacturer's instructions. First, DNA was sheared to 350 bp fragments using the Covaris S220 system and then cleaned up using Sample Purification Beads (Illumina). 10 μ l CTE and 40 μ l ERP2 were added to a final volume of 100 μ l and then incubated at 30 °C for 30 min to convert overhangs. Purification beads were used to perform right- and then left-sided size selection, and DNA was eluted in 15 μ l resuspension buffer (RSB). 3' ends were adenylated by adding 2.5 μ l CTA and 12.5 μ l ATL and then incubated at 37 °C for 30 min and then 70 °C for 5 min. Then, 2.5 μ l CTL, 2.5 μ l LIG2,

and 2.5 μ l unique DNA adapters were added, mixed, incubated at 30°C for 10 min, and finally 5 μ l STL was added. Libraries were cleaned up using purification beads and eluted in 50 μ l RSB.

Library fragment size was determined using an Agilent Bioanalyzer 2100 (DNA 1000) and quantified by real-time PCR using the Kapa library quantification kit (Kapa Biosystems) on the QuantStudio 6 Flex (Applied Biosystems). Libraries were then diluted to 0.75 nM and pooled in 12-plex according to the D7 indexes (columns). Pooled libraries were sequenced on an Illumina X Ten according to manufacturer's instructions to produce paired-end 150 bp reads. Each pool of 12 libraries was sequenced over eight lanes.

2.12.1 Computational analyses of whole genome sequencing

Sequencing read alignment

Sequencing reads were aligned to the GRCm38 mouse genome assembly (Ensembl release 91 (Yates et al., 2016)) with bwa (versions 0.6.1 or 0.7.12 (Li and Durbin, 2009)). Aligned reads were annotated to read groups using the picard tool AddOrReplaceReadGroups and minor annotation inconsistencies corrected using the picard CleanSam and FixMateInformation tools (picard version 1.124; <http://broadinstitute.github.io/picard>). The bam files for each sample were merged together and duplicate reads were then identified using the picard MarkDuplicates tool. Sequencing coverage was assessed using samtools (version 1.1 (Li et al., 2009)).

Aligned reads for human samples from the ICGC French liver cancer (LICA-FR) cohort were downloaded from the European Genome-phenome Archive at EMBL-EBI (accessions EGAD00001000131, EGAD00001001096 and EGAD00001000737).

Variant calling

A pooled normal sample set was generated by combining all control sample reads and then sub-sampling these reads to match the mean coverage achieved for the control and tumour samples. Single nucleotide and indel variants were called using Strelka (version 1.0.14 (Saunders et al., 2012)) using the recommended configuration for bwa-aligned reads and setting the `isSkipDepthFilters` flag for improved calling. Variant calls were combined into a merged set using bcftools (version 1.1 (Li et al.,

2009)). Predicted coding sequence changes due to the SNVs were annotated by comparing the polypeptide sequences coded by the reference and variant alleles.

SNVs were subjected to multiple filtering steps. First, low-confidence SNV calls were identified and removed by applying the recommended filters for Strelka output from the gatk-tools package (version 0.2 (Alioto et al., 2015); <https://github.com/crukci-bioinformatics/gatk-tools>). In particular, variants were filtered based on low mapping and base quality scores, proximity to alignment ends, and low absolute read counts. Secondly, SNVs with an allele frequency of less than 2.5% were omitted to eliminate possible cross-contamination due to observed low levels of sequence read index mis-assignment during Illumina sequencing (Owens et al., 2018).

SNV call rates were estimated by fitting the number of variants detected at a range of sequencing read depths and extrapolating to determine the expected call rate at saturated sequencing coverage. Variant allele frequencies were calculated from read counts for reference and variant alleles, excluding those reads having a MAPQ score less than 5. Confidence intervals for allele frequency estimation were estimated by applying a normal distribution approximation to bootstrap-resampled frequency estimates (R boot package version 1.3 (Davison and Hinkley, 1997)). The SVN data are deposited at the European Nucleotide Archive under accession code ERZ537503.

Autosomal copy number variations (CNVs) were called with CNVkit (version 0.7.2 (Talevich et al., 2016)) using default parameters. CNV regions were filtered to remove low-confidence regions where the null hypothesis (i.e., unchanged copy number) fell within the 95% confidence interval. Further filters removed CNVs where the absolute log fold-change in copy number was smaller than 0.25. CNV regions located within 10 kb of each other were merged, and the resulting regions finally filtered to remove regions smaller than 10 Mb.

SNV validation

Non-synonymous SNVs in the cancer driver genes of interest, *Braf*, *Hras*, and *Apc*, were validated using conventional Sanger sequencing. In addition, SNVs were checked by visual inspection of the aligned reads and were called validated if the total variant reads were greater than ten.

Mutational signature analysis

Analysis of mutational signatures was constrained to just those regions covered to at least 20x in all samples. The distributions of 5' and 3' nucleotides flanking the called SNVs were calculated directly from the reference genome. Direct comparison between human and mouse signatures was facilitated by normalising C57BL/6 nucleotide context distributions using the ratios of known trinucleotide prevalences in mouse and human genomes, as calculated for the 20x covered regions for each genome. The proportions of COSMIC mutational signatures (Forbes et al., 2017) represented in the mutational profile from each sample were calculated using the R package *deconstructSigs* (version 1.8.0 (Rosenthal et al., 2016)).

Identification of significantly mutated cancer-related genes

Variants were annotated as to their likely effect on coding sequence by comparing their predicted polypeptide sequences to those from the Ensembl release 91 GRCm38.p5 reference. SNV calls shared between tumours taken from the same mouse (i.e., SNVs that could be simply ascribed to germline variation) were filtered prior to analysis of mutated genes.

Cancer-related genes bearing above expected levels of non-synonymous mutations (both across the entire gene and recurring at specific loci) were identified using the following procedure. The gene list to be analysed was constrained to the listing of oncogenes and tumour suppressor genes described by Vogelstein et al. (2013). The total mutation load (nonsense, missense, frame shift, splice site) within coding and splice-site regions was used to calculate the probability that mutations had occurred purely by stochastic mutational processes. Within each gene, the count of mutations at each variant locus was fitted to a Poisson distribution assuming a background mutation rate calculated across all sequenced regions. The individual variant loci were combined at the gene level using a multinomial model using the R *XNomial* package (version 1.0.4; <https://CRAN.R-project.org/package=XNomial>). This model yielded log-likelihood ratios from the observed and expected distributions, from which a gene-wise p-value was readily calculated. P-values were corrected for multiple testing across all genes using the Bonferroni method. The analysis was repeated imposing a gene filter derived from the COSMIC Cancer Gene Census (<http://cancer.sanger.ac.uk/census/>). This identified the same cancer-associated genes that were significantly mutated in the mouse neoplasms.

Mutational enrichment analysis

Mutation rate estimation, background mutation rates, and enrichment analyses were performed as Sabarinathan et al. (2016). Briefly, mutation rate was estimated for all CTCF binding sites identified from ChIP-seq experiments plus flanking stretches of 400 nucleotides at both sides. To avoid bias, windows of 801 nt were excluded if they contained coding sequences or UCSC Browser blacklisted regions (“CRG Alignability 36’ Track”, score < 1, <http://genome.ucsc.edu/cgi-bin/hgFileUi?db= hg19&g= wgEncodeMapability>). The resulting filtered windows were aligned taking CTCF as the centre, and the mutation rate of every column i within the window was calculated as the total number of mutations mapped to nucleotides in column i divided by the total number of nucleotides observed in column i after filtering.

In order to check if the mutation rate observed at each position was expected due to the local sequence context, we randomly introduced the same number of mutations observed at each window following the probability of occurrence of each mutation according to its trinucleotide context. The mutation rate of each randomly generated set of changes was computed for each column as explained above. This procedure was repeated 1,000 times to compute the mean random mutation rate of every column in the motif.

To test the enrichment for mutations at CTCF binding sites compared to their flanking regions, we compared the ratio of the total number of mutations to the total number of nucleotide positions within the motif (± 15 nt) and that of the flanking region (16 to 400 nt) on either side using a chi-squared test. In addition, we computed the fold change of mutation rates through the expected frequencies obtained from chi-squared tests. P-values were corrected for multiple-testing using the Benjamini–Hochberg procedure.

2.13 Data storage and management

All data were stored on secure servers at the CRUK Cambridge Institute, and access to the repository was restricted to Odom group members. In addition, access was shared on a per-project basis with external collaborators at EMBL-EBI and IRB Barcelona. All data generated are managed and shared in compliance with the Data Management Policy at Cambridge University (<http://www.lib.cam.ac.uk/dataman/pages/bidding.html>), which is a strong advocate of open access of research data.

The raw data associated with peer-reviewed publications are deposited in the ArrayExpress repository under accession codes E-MTAB-6261 for the ChIP-seq dataset (including peak calls), E-MTAB-6259 for the RNA-seq dataset (including raw and normalised gene expression counts), E-MTAB-6262 for the Hi-C dataset, and E-MTAB-6971 and E-MTAB-6972 for the mouse liver tumour and control datasets (including raw and normalised RNA-seq counts). The SVN data for BL6 mouse derived tumours are deposited at the European Nucleotide Archive under accession code ERZ537503. H&E-stained tissue section images of the tumours included in whole genome sequencing experiments are available at BioStudies archive at EMBL-EBI under accession S-BSST129.

Chapter 3

CTCF maintains regulatory homeostasis of cancer pathways

Since thesis submission, this chapter has been published in full as Aitken et al. (2018) *Genome Biology*.

3.1 Introduction

As detailed in Chapter 1, CTCF is a highly conserved nuclear phosphoprotein (Klenova et al., 1993; Ohlsson et al., 2001) that is ubiquitously expressed in somatic cells (Phillips and Corces, 2009). It is responsible for diverse regulatory functions including fine-tuning gene expression, X chromosome inactivation, imprinting, and 3D chromatin organisation (Bell et al., 1999; Bickmore, 2013; Fedoriw et al., 2004; Filippova et al., 1996; Lieberman-Aiden et al., 2009; Tsai et al., 2008). The global three-dimensional (3D) organisation of chromatin partitions the mammalian genome into discrete structural and regulatory domains (Dixon et al., 2012; Lieberman-Aiden et al., 2009). Chromosome architecture has multiple levels of spatial organisation: megabase-scale compartments correspond to euchromatin (A) and heterochromatin (B) (Simonis et al., 2006), sub-megabase regions can be defined as topologically associated domains (TADs) (Dixon et al., 2012) and, at the tens of kilobases level, there exist smaller loop structures that connect *cis*-regulatory elements (Rao et al., 2017, 2014). CTCF is frequently present at these structural boundaries at all scales (Moore et al., 2015; Vietri Rudan et al., 2015).

Numerous studies have explored the function of complete disruption of CTCF binding both *in vivo* and *in vitro*. At the whole embryo level, homozygous *Ctcf*

deletion is embryonic lethal (Fedoriw et al., 2004), and genetically inducible *Ctcf* knockout in specific cell types including oocytes (Wan et al., 2008), lymphocytes (Ribeiro de Almeida et al., 2011), neurons (Hirayama et al., 2012), and cardiomyocytes (Lee et al., 2017) results in organ-specific failure, characterised by aberrant enhancer-promoter interactions and transcriptional dysregulation (Downen et al., 2014). Complementary biochemical approaches have been used to test the functional impact of acute CTCF depletion *in vitro* by both RNAi (Schmidt et al., 2012; Zuin et al., 2014) and transient auxin-mediated depletion (Nora et al., 2017). Acute depletion in mouse embryonic stem cells results in almost complete removal of CTCF from the nucleus, causing genome-wide disruption of loops and TADs without affecting higher-order genomic compartmentalisation (Nora et al., 2017).

Despite strong conservation of the higher-order chromatin structure, such as TADs, across tissues and individuals (Vietri Rudan et al., 2015), there is substantial inter- and intra-individual variation in *Ctcf* expression (Phillips and Corces, 2009) driven by both genetic heterogeneity and cell type specificity. Up to ten-fold differences in both *Ctcf* mRNA and protein expression have been observed across a variety of tissues (Mele et al., 2015; Uhlen et al., 2015). Since these differences in expression do not seem to affect the general organisation of chromatin, it is unclear whether they have a functional impact. To address this, we sought to develop and exploit a highly controlled system in which we could modulate *Ctcf* expression without resorting to a homozygous knockout.

We therefore utilised mice with hemizygous *Ctcf* deletion (Heath et al., 2008; Lallemand et al., 1998) in an attempt to dissect direct regulatory targets and function (Boj et al., 2010). While *Ctcf* hemizygous mice develop normally, they have are predisposed to tumour development (Kemp et al., 2014), suggesting that even physiologically-tolerated changes in CTCF concentration have a detrimental effect on organismal fitness. *CTCF* is also implicated as a haploinsufficient tumour suppressor gene in human cancers (Filippova et al., 1998; Kemp et al., 2014; Ohlsson et al., 2001).

In contrast to germline variants, somatic missense and nonsense mutations of *CTCF* are common in human cancers (Gonzalez-Perez et al., 2013; Rubio-Perez et al., 2015). *CTCF* has been identified as a putative driver gene in several cancer types (Rubio-Perez et al., 2015) and such loss of function is in keeping with the action of a tumour suppressor gene (Filippova et al., 1998; Ohlsson et al., 2001). Furthermore, reduced expression of *CTCF* mRNA in kidney cancer is strongly correlated with lower five-year survival rates (Uhlen et al., 2017). However, the

precise role of *CTCF* in the initiation or progression of carcinogenesis is poorly understood.

To study the direct impact of altering *Ctcf* expression, independent of any factors that may confound human studies such as environmental exposures, we chose an *in vitro* model and exploited mouse embryonic fibroblasts (MEFs). Wild-type and *Ctcf* hemizygous mouse embryonic fibroblasts MEFs were interrogated using a variety of functional assays to characterise differences in the molecular portraits between conditions. MEFs were chosen in these studies because: (i) they are grown in well-defined culture conditions *in vitro* and (ii) their rapid growth rate allowed us to harvest the volume of cells needed for each ChIP-seq, RNA-seq, proteomic, and Hi-C replicate from a single embryo at a low passage number (P4). The low passage number was important because extended passage runs the risk of the introduction of mutations and transformation.

3.1.1 Project aim and overview

The aim of this project was to investigate the effects of *Ctcf* haploinsufficiency by generating, analysing, and integrating genomic, epigenomic, transcriptomic, proteomic, and chromatin-conformation data in a mouse model of *Ctcf* hemizyosity (Figure 3.1).

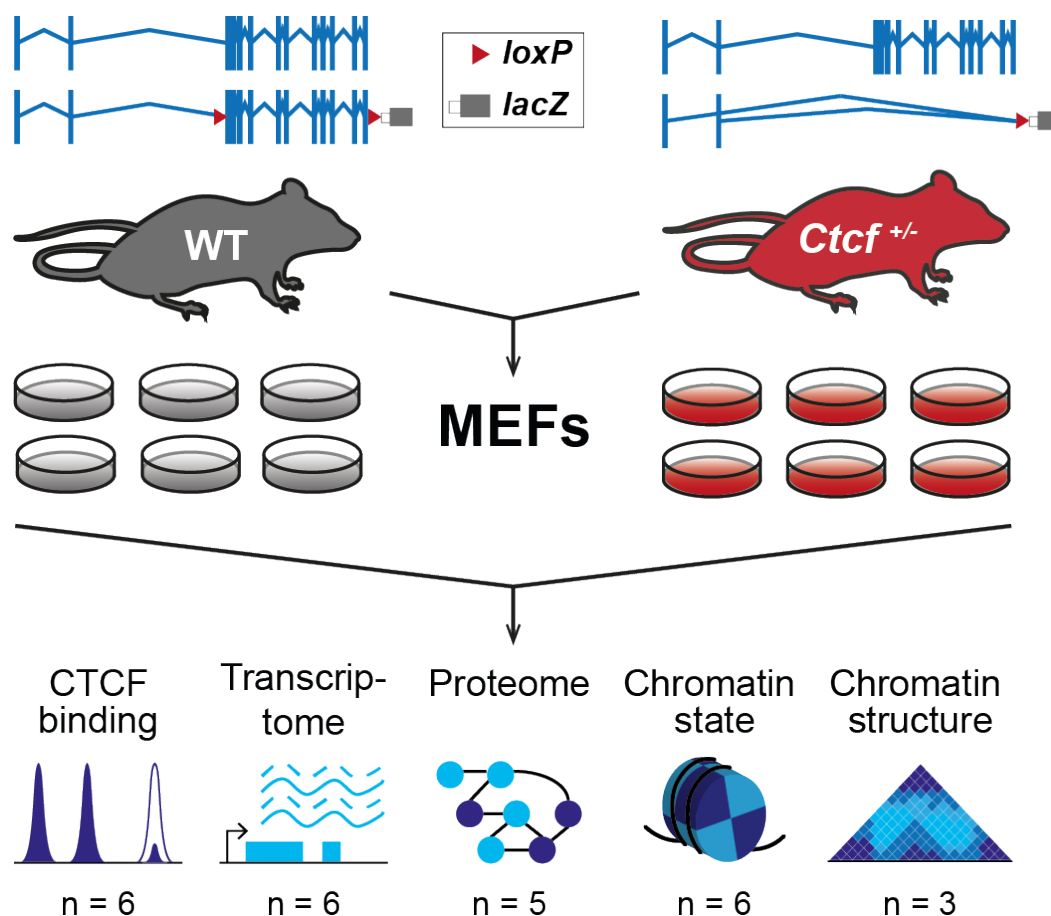


Fig. 3.1 Project overview: *Ctcf* hemizyosity as a model to subtly perturb nuclear homeostasis. The engineered *Ctcf* locus contains *loxP* sites flanking the protein-coding exons of the gene (WT, *Ctcf*^{+/+}), which can be removed using Cre recombinase (*Ctcf*^{+/-}). Mouse embryonic fibroblast (MEF) lines were derived from six WT and six *Ctcf*^{+/-} littermates. Quantitative analyses of CTCF binding, transcription, proteomics, chromatin state, and chromatin structure were performed in multiple biological replicates.

3.2 Results

3.2.1 Successful generation of embryonic fibroblast cultures

To characterise the molecular effects of altering the concentration of CTCF protein available in the nucleus, we utilised *Ctcf* hemizygous mice carrying a *lacZ* reporter in place of the coding region of *Ctcf* (Heath et al., 2008) in all cells (**Figure 3.2**). To perform high-throughput experiments in homogeneous cell populations, an *in vitro* approach was employed.

Wild-type *Ctcf* allele

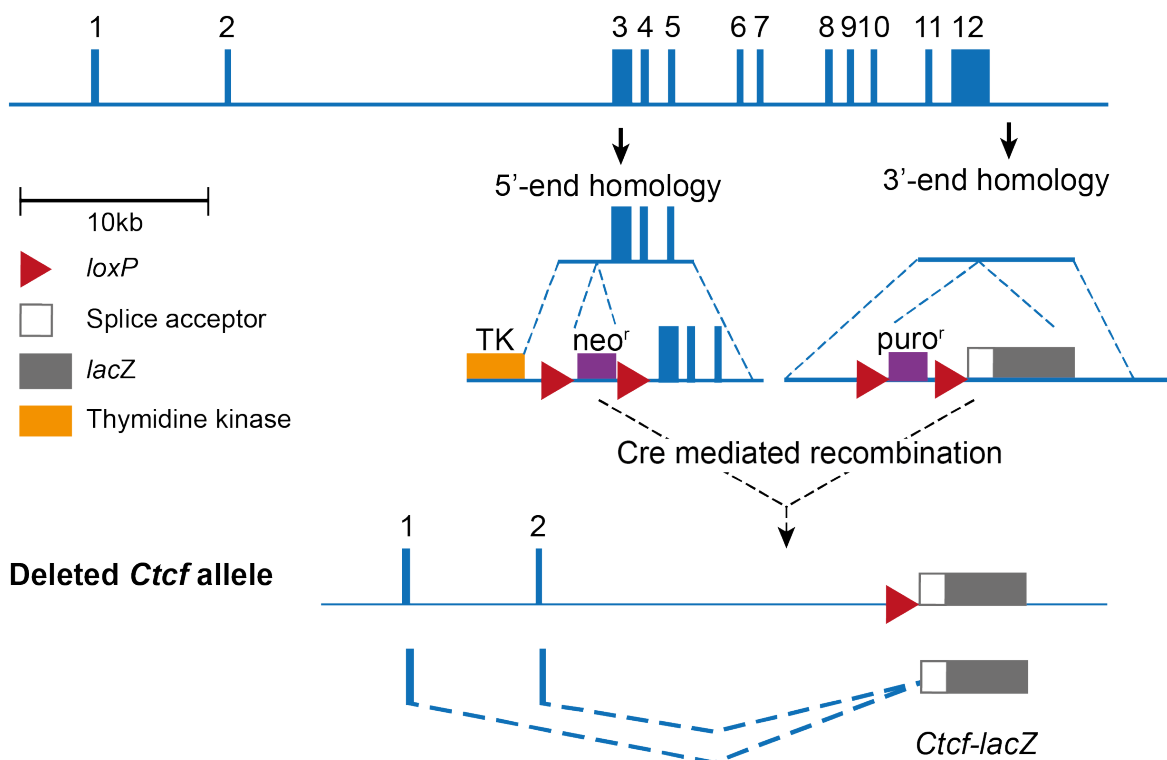


Fig. 3.2 Conditional deletion of the mouse *Ctcf* gene. Exons of the wild-type *Ctcf* gene (solid boxes) are numbered (10 kb scale indicated). Exon 3 contains the start codon and exon 12 contains the stop codon. The two targeting constructs with *loxP* sites (red triangles) flanking a PMC1-neomycin cassette (neo^r) or a PGK-puromycin cassette (puro^r) are shown with homologous regions (Hoogenraad et al., 2002). Complete Cre-mediated recombination at the outermost *loxP* sites deletes the coding exons (3-12) of the *Ctcf* gene, shown in the lower panel. Alternative splicing generates a hybrid *Ctcf-lacZ* transcript by splicing of the splice acceptor site present at the 5'-end of the reporter *LacZ* cassette to *Ctcf* exon 1 or 2.

Timed mouse matings were used and, after identification of a vaginal plug on day E0.5, embryos were collected on day E13.5 and MEF cultures generated. Embryos were harvested from three pregnant females at E13.5 to produce 21 MEF cultures. PCR analysis of embryos confirmed both wild-type (WT, *Ctcf*^{+/+}) and *Ctcf* hemizygous (HE, *Ctcf*^{+/-}) genotypes (**Table 3.1**). All cells cultured successfully, reaching 80% confluence by 72 h after splitting (**Figure 3.3**). MEFs were cultured and cells were harvested in pairs (one WT and one *Ctcf* hemizygous) to account for potential batch effects cause by culture conditions.

Table 3.1 Genotyping of mouse embryos

Maternal Mouse ID	Wild-type <i>Ctcf</i> ^{+/+}	Hemizygous <i>Ctcf</i> ^{+/-}
AN15CUK011169	3	4
AN15CUK015753	0	5
AN15CUK018389	3	6
Total no. embryos	6	15

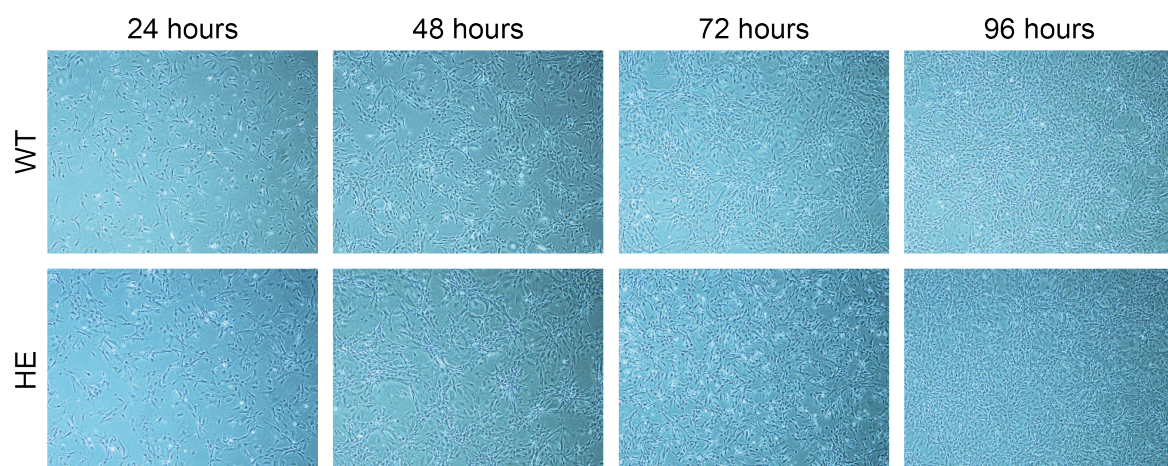


Fig. 3.3 MEF cultures. Photomicrographs of wild-type (WT) and *Ctcf* hemizygous (HE) MEF cultures 24, 48, 72, and 96 hours after first passage. Original magnification x10.

3.2.2 Molecular characterisation of *Ctcf* hemizygous MEFs

We used six of our independently derived embryonic fibroblast lines from mice carrying a deletion of one *Ctcf* allele (*Ctcf*^{+/-}) and six corresponding lines from littermate *Ctcf* wild-type controls. Each MEF culture showed significant depletion of *Ctcf* mRNA (**Figure 3.4A**) and a slightly lesser reduction in CTCF protein (**Figure 3.4B**); the degree of *Ctcf* depletion varied between biological replicates. qPCR demonstrated that, on average, *Ctcf* hemizygous MEFs had a 37% reduction (two-tailed t-test, $p = 1.58 \times 10^{-6}$) in *Ctcf* mRNA compared to wild-type MEFs. In turn, quantitative western blotting showed a 27% reduction (two-tailed t-test, $p = 8.731 \times 10^{-5}$) in protein level versus wild-type cells (**Figure 3.4C**). Thus, although there is partial compensation at both the mRNA and protein levels, there is a consistently lower concentration of CTCF in hemizygous mouse cells.

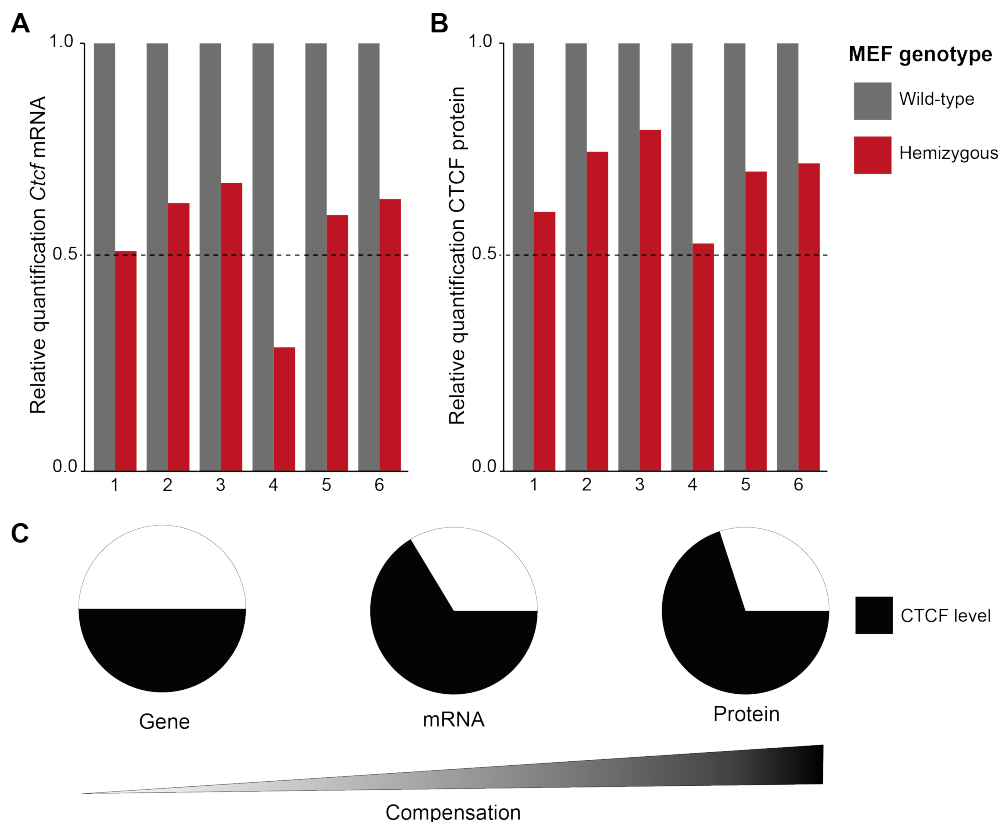


Fig. 3.4 Validation of CTCF depletion. Quantification of *Ctcf* deletion by (A) qRT-PCR and (B) quantitative western blot experiments on wild-type (*Ctcf*^{+/+}) and hemizygous (*Ctcf*^{+/-}) MEFs. (C) There is partial compensation from DNA to RNA to protein levels of CTCF.

Given this haploinsufficient molecular phenotype, we anticipated that this *Ctcf* hemizygous model system would provide a powerful tool to delineate the pathophysiological consequences of CTCF haploinsufficiency. We therefore used these twelve independent embryonic fibroblast lines to generate an adequate volume of cells for multiple biological replicates for diverse functional experiments (**Figure 3.1**).

3.2.3 Chronic reduction of CTCF alters its chromatin binding

We first assessed the impact of hemizyosity on CTCF occupancy using chromatin immunoprecipitation followed by sequencing (ChIP-seq). We identified 42,336 loci directly occupied by CTCF, 787 of which were significantly differentially bound (FDR <5%) between the two genotypes (**Figure 3.5A**). Of these, 79.4% were less strongly bound in the *Ctcf*^{+/-} MEFs. The changes in occupancy between the genotypes were generally small but highly reproducible among independent samples (**Figure 3.5B**). Thus, reduced availability of CTCF in embryonic fibroblasts leads to its depletion at a very specific subset of genomic sites.

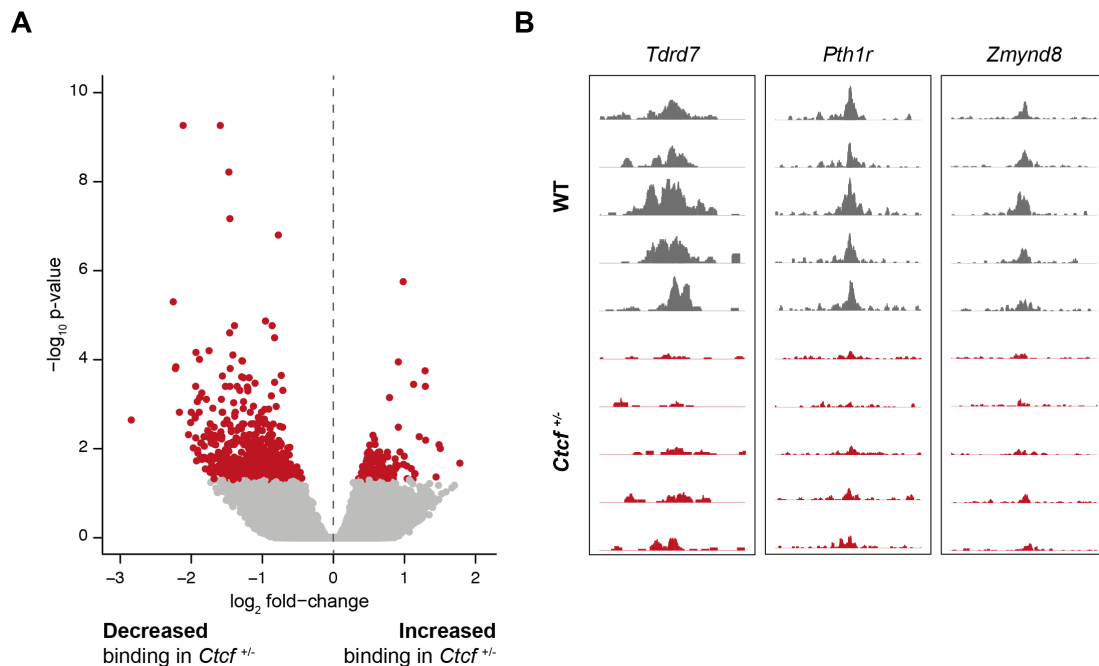


Fig. 3.5 *Ctcf* hemizyosity results in altered chromatin binding. (A) Differential binding analysis identified 787 CTCF binding sites differentially occupied between *Ctcf* hemizygous and wild-type MEFs, most of which show reduced genomic occupancy in the *Ctcf*^{+/-} MEFs. Significant changes are shown in red (FDR < 5%). (B) Example genome tracks showing highly consistent loss of CTCF binding at three genomic loci overlapping the genes indicated at the top of the panels. The same scale is used on all y-axes.

3.2.4 Labile CTCF binding sites have distinct genomic features

Genomic locations sensitive to subtle and chronic CTCF reduction shared a number of features. First, most (68%) of the differentially bound sites overlapped annotated genes or their promoters (defined as 5 kb upstream of the transcription start site), which represents significant enrichment compared to genome-wide CTCF occupancy (chi-square test, $p = 4.9 \times 10^{-10}$; **Figure 3.6A**). Second, CTCF can bind motif instances of either 20 or 33 bases (Schmidt et al., 2012); we found that differentially bound CTCF sites were significantly depleted of longer words (hypergeometric test, $p = 1.53 \times 10^{-11}$; **Figure 3.6B**). Previous studies have shown that binding sites with shorter motifs have lower average binding affinity (Schmidt et al., 2012). Consistent with this, CTCF sites perturbed by hemizyosity had motifs of lower affinity when compared to all CTCF-bound regions (Mann-Whitney test, $p < 2.2 \times 10^{-16}$; **Figure 3.6C**). Third, by comparing with the ~11,000 CTCF sites conserved across five species of mice (Thybert et al., 2018), differentially-bound CTCF sites were found to be depleted of these conserved binding events (hypergeometric test, 2.55×10^{-6} ; **Figure 3.6D**). In other words, CTCF binding sites stable across the murine lineage are resistant to chronically reduced levels of CTCF.

In conclusion, our data reveal a set of regions preferentially found near genes which show reproducible, quantitative changes in CTCF occupancy that show common motif characteristics and that are enriched for lineage-specific CTCF binding.

3.2.5 *Ctcf* hemizyosity alters transcription of cancer pathways

To determine what impact changes in CTCF binding had on the global transcriptome, we sequenced total RNA from six biological replicates of both genotypes. Confirming the qPCR results, hemizyosity resulted in a significant reduction in *Ctcf* expression ($p = 2.4 \times 10^{-7}$, Methods). Consistent with the differences in CTCF occupancy, transcriptional changes were subtle: differential gene expression analysis identified 296 dysregulated genes (FDR <5%; **Appendix A**) 69% of which had reduced expression in *Ctcf*^{+/-} MEFs (**Figure 3.7A**).

mRNA and cellular protein levels have a complex relationship (Liu et al., 2016). In our model, the changes observed in the transcriptome did propagate to the protein level, as shown by comparison of transcriptomic data to the proteomes of the wild-type and hemizygous cells obtained using tandem mass tag (TMT) proteomics. For the differentially expressed genes, the transcriptional and proteomic changes in *Ctcf* hemizygous cells were highly correlated (Spearman's $\rho = 0.65$, $p < 2.2 \times 10^{-16}$;

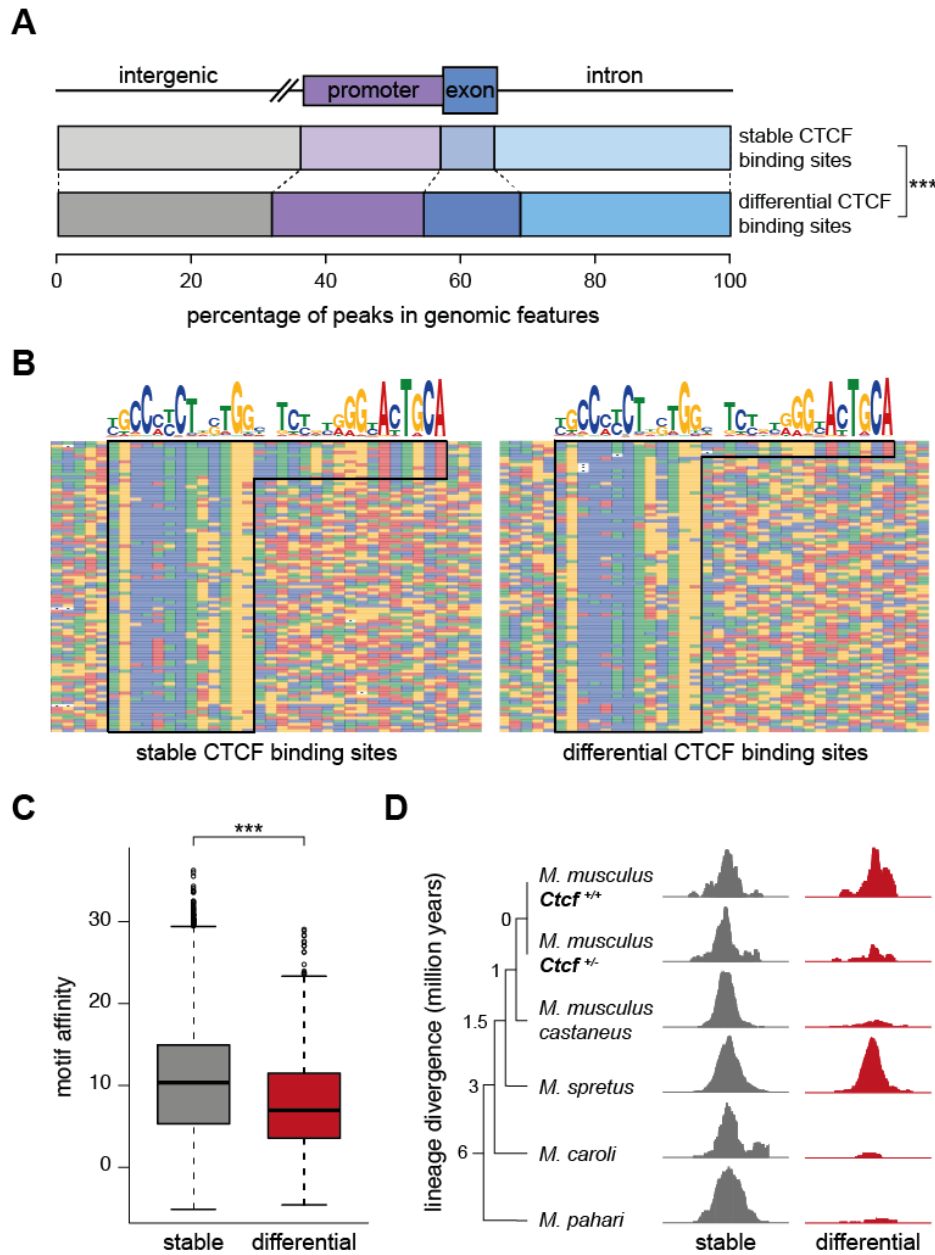


Fig. 3.6 Differentially bound CTCF loci are found near genes, occur in shorter motifs, and have lower binding affinity and evolutionary conservation. (A) Differential CTCF binding sites were significantly enriched within promoters and gene bodies compared to stable CTCF binding sites (chi-square test, $p = 4.9 \times 10^{-10}$). (B) Stable CTCF peaks had a higher proportion of the longer (~33 bp) motif word compared to the differential sites. Multiple alignments of a randomly chosen subset of a hundred CTCF binding sites that are either stable or differential are shown. Each position in the alignment is coloured corresponding to the nucleotide present following the colour scheme used in the logo shown at the top. (C) Binding sites susceptible to a reduced CTCF concentration have significantly lower motif affinity (Mann-Whitney test, $p < 2.2 \times 10^{-16}$). (D) Regions bound by CTCF across the mouse lineage are less sensitive to *Ctcf* hemizyosity. Example tracks are shown of a stable CTCF binding site that is conserved in five species of mice compared to a differential site that is found in only a subset of the species (*M. musculus* chr6:120,736,800 for the stable site and chr2:31,887,060 for the differential site). *** p value < 0.001 .

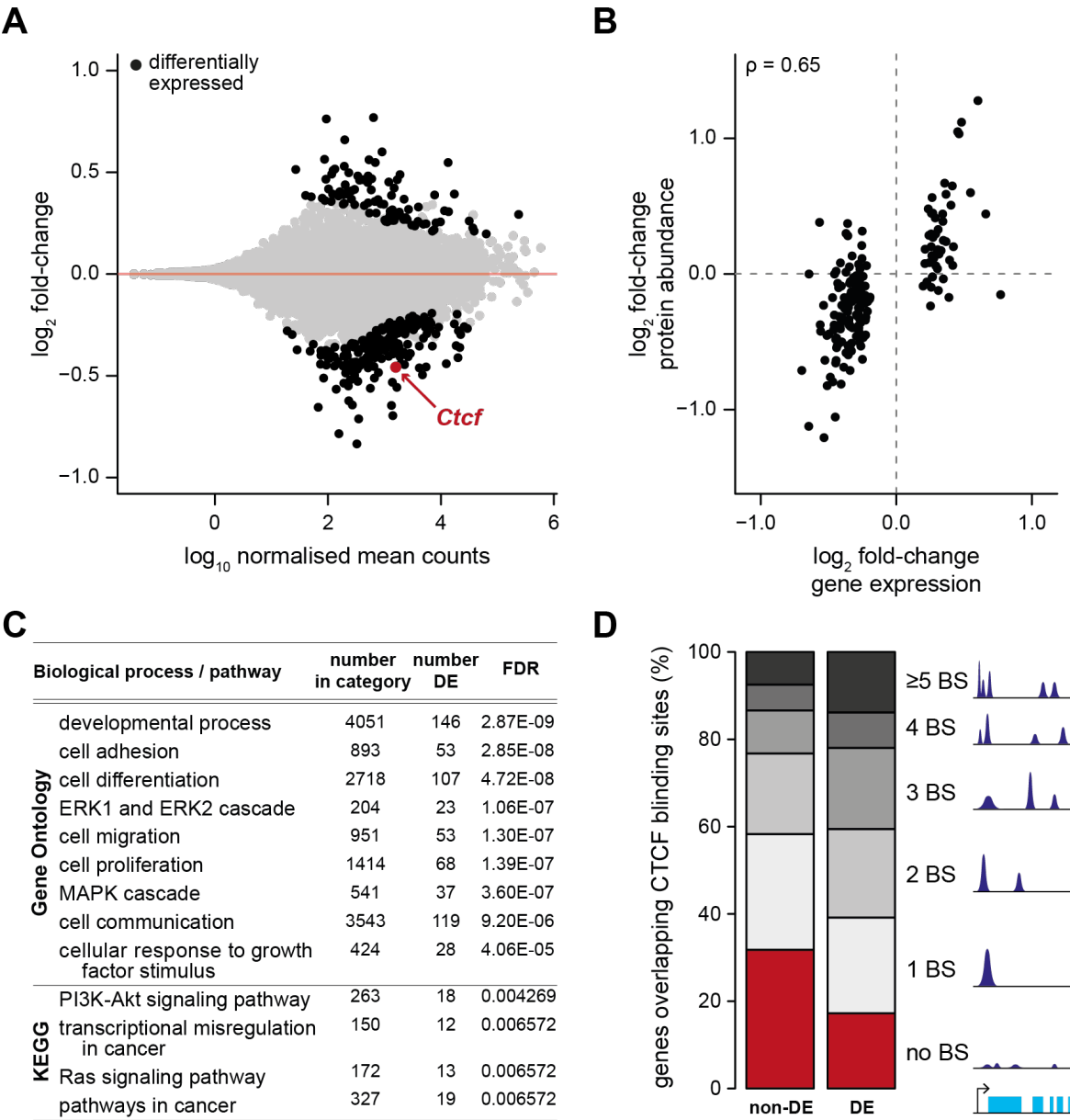


Fig. 3.7 CTCF depletion dysregulates oncogenic pathways. (A) Nearly 300 genes were differentially expressed between wild-type and *Ctcf*^{+/-} cells; significant changes are shown in black (FDR <5%). (B) Transcriptional changes (x-axis) were highly correlated to corresponding changes at the protein level (y-axis); Spearman's correlation coefficient is noted in the top left corner. 86% of all genes from (A) had concordant fold-change estimates in the proteomics dataset. (C) Gene set enrichment analysis performed on the differentially expressed genes highlights dysregulation of cancer-related pathways. Representative significantly enriched terms from the Gene Ontology (upper panel) and KEGG pathways (lower panel) are shown. (D) Differentially expressed (DE) genes are strongly enriched for having higher numbers of CTCF binding sites (BS) than genes with stable expression, in their gene bodies or flanking 5 kb.

Figure 3.7B). Indeed, 85% had fold-change estimates concordant between RNA transcription and protein expression.

Gene set enrichment analysis revealed that CTCF-dependent transcripts were strongly enriched for processes related to cell differentiation, proliferation, death, migration, adhesion, angiogenesis, and protein phosphorylation. The MAPK, ERK1/2 and Ras signalling pathways also showed an excess of dysregulated transcripts. Consistent with these results, analysis of KEGG pathways revealed that *Ctcf* hemizygosity resulted in perturbation of cancer-related pathways (**Figure 3.7C**).

Finally, we asked whether these gene expression changes could be caused directly by altered CTCF binding. We observed that few differentially bound CTCF sites overlapped, or were in close proximity to, the genes with altered expression. However, there was a strong tendency for differentially expressed genes to be associated with higher numbers of CTCF binding sites (**Figure 3.7D**), even if these were stable. For example, 83% of all dysregulated genes (± 5 kb) overlapped at least one CTCF binding site, in contrast to only 68% of stable genes (hypergeometric test, $p = 1.32 \times 10^{-8}$). Further, whereas only 23% of stable genes overlapped with three or more CTCF bound sites each, 41% of all differentially expressed genes did (**Figure 3.7D**). Thus, the set of genes dysregulated in *Ctcf* hemizygous cells are strongly enriched for CTCF binding sites, suggesting subtle additive effects regulate nearby gene transcription.

3.2.6 Gene expression changes correspond with altered looping interactions

Steady-state transcription can be altered by either changes in transcript stability or by differences in transcriptional regulation. We therefore examined whether the promoters of CTCF-dependent genes showed corresponding changes in transcriptional initiation, reflected as changes in H3K4me3 and H3K27ac occupancy. Both of these histone modifications are associated with an open chromatin state, are permissive of active transcription (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Santos-Rosa et al., 2002; Shlyueva et al., 2014), and their occupancy levels at the transcription start site are positively correlated with gene expression (Santos-Rosa et al., 2002). The vast majority ($>80\%$) of the CTCF-dependent genes had concordant promoter and transcriptional changes (**Figure 3.8**). Thus, most gene expression differences apparently arise from CTCF-mediated alterations to transcriptional initiation.

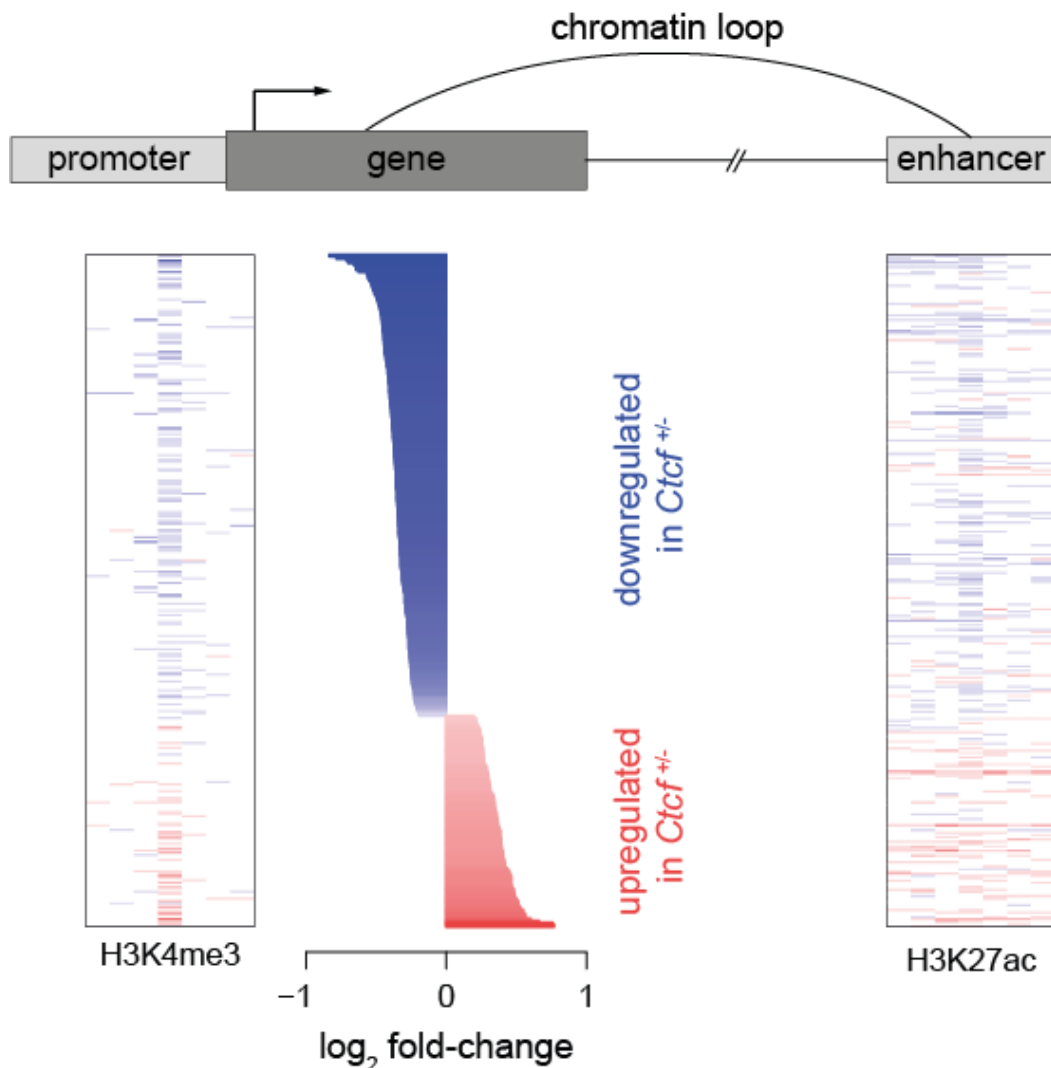


Fig. 3.8 Transcriptional perturbations arise from regulatory changes in the nuclear genome. Changes in expression of dysregulated genes are accompanied by changes in the activity of their proximal promoters as well as enhancers linked via chromatin loops. In the centre, the expression differences between wild-type and *Ctcf*^{+/-} cells are shown, ordered by increasing fold-change; genes expressed at lower levels in the hemizygous cells are in blue, whereas those expressed at higher levels are in red. To the left, a heatmap of the difference in mean abundance of H3K4me3 occupancy is shown. Each column is a 5 kb window extending 17.5 kb up- and downstream of each gene's transcription start site, which is in the centre. On the right, an equivalent heatmap for the difference in occupancy of H3K27ac, centred at the midpoint of the peak. Gene-enhancer pairs were inferred from significant interactions identified from Hi-C data, and thus elements can be separated by large distances. For each gene, the enhancer with most regulatory potential is shown (Methods). The same colour scale is used throughout. Transcriptional changes are accompanied by concordant changes in the activity of their regulatory elements.

Gene expression can be controlled by looping interactions between regulatory elements mediated by CTCF (Bonn et al., 2012; Chambeyron and Bickmore, 2004; Fraser, 2006; Ling et al., 2006; Splinter et al., 2006; Yoon et al., 2007). To determine the effects of *Ctcf* hemizyosity on chromatin architecture, nuclear Hi-C experiments were performed in three biological replicates of wild-type and *Ctcf*^{+/-} MEFs. First, we inspected the global-scale interaction profiles in both genotypes using 100 kb windows covering the whole genome. Consistent with recent studies showing that acute total depletion of CTCF results only in modest effects on large-scale chromatin interactions (Nora et al., 2017), we found that 95% of all windows were unaffected by reduced CTCF, with a correlation coefficient of 0.9 or higher between the genotypes (**Figure 3.9**).

To explore whether fine-scale chromatin organisation is affected by reduced CTCF levels, we merged all replicates to increase the resolution of our data and identified pairs of loci interacting more often than expected by chance. We then compared the intensity of such interactions in the wild-type and hemizygous cells and generated a ranked list. Looping interactions that involved a dysregulated gene or a differentially bound CTCF site tended to rank higher and were significantly enriched at the top of the list (Wilcoxon signed rank test, $p = 0.017$ for genes and $p = 2.72 \times 10^{-6}$ for CTCF sites). Thus, many of the transcriptional changes we observed may indeed be the result of changes in distal regulatory elements mediated by looping interactions.

We reasoned that these loops were likely to connect altered genes to distal enhancer activity changes. We defined putative enhancers as sites occupied by H3K27ac but lacking H3K4me3 (Heintzman et al., 2007; Rada-Iglesias et al., 2011) and identified 73,670 loci with this epigenetic profile. We then collected the subset of enhancers associated with a dysregulated gene via a looping interaction (Methods), and compared the fold-change between the wild-type and *Ctcf* hemizygous cells. 75% of these enhancer-gene pairs showed concordant changes between gene expression and enhancer activity (**Figure 3.8**). Bulk analysis of enhancer changes would not have identified these connections, since direct comparison of hemizygous and wild-type *Ctcf* cells showed almost no enhancer differences, with only 127 (0.2%) being significantly differentially bound ($\text{FDR} < 5\%$). Thus, the transcriptional changes observed in the hemizygous cells are likely to result from altered transcriptional regulation mechanisms that involve both promoters and distal enhancers.

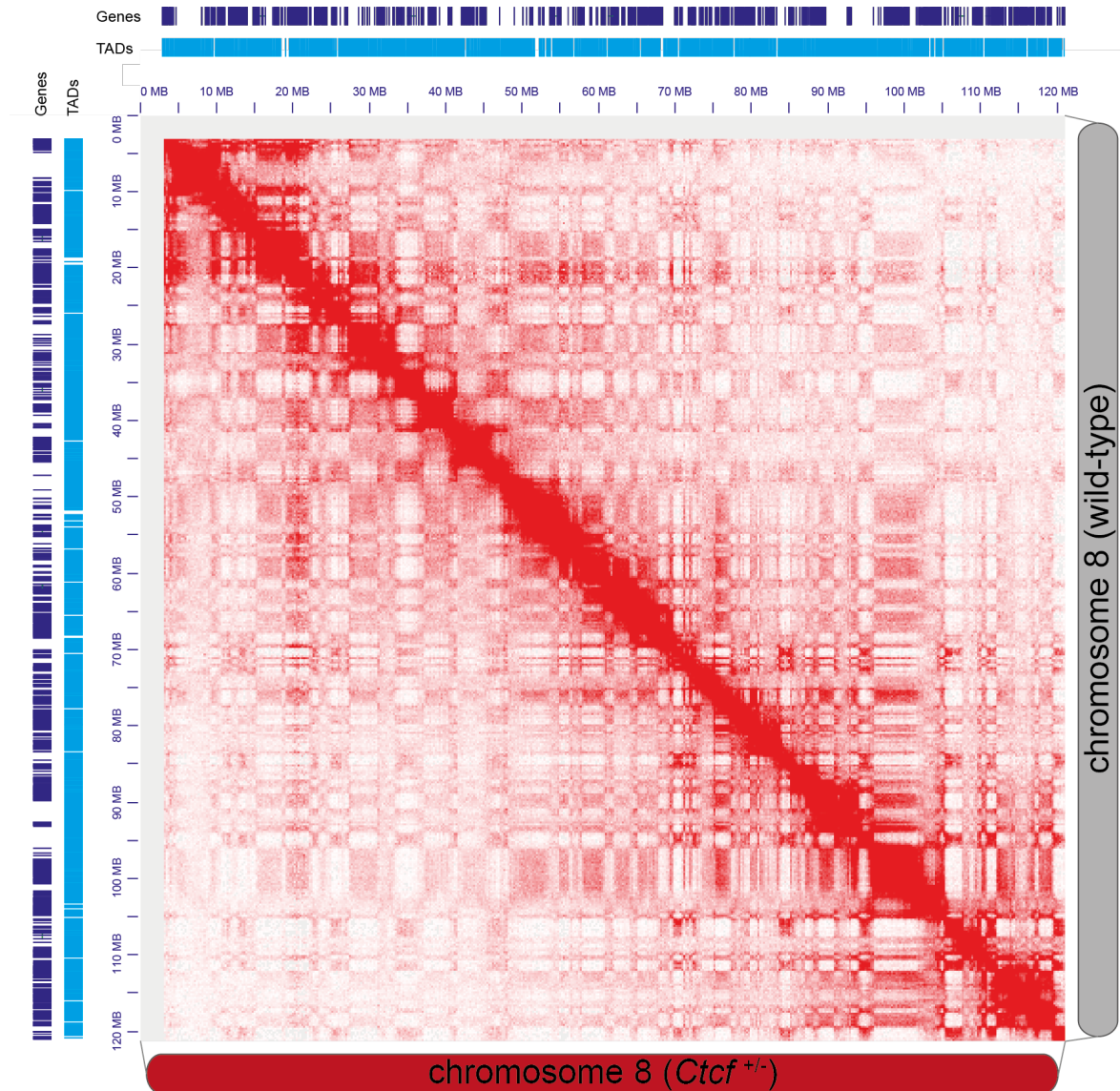


Fig. 3.9 Global-scale chromatin interactions are robust to reduced CTCF levels. Hi-C analysis produces a genome-wide contact matrix; the submatrix shown here corresponds to intrachromosomal interactions on chromosome 8. Each pixel represents all interactions between one 100 kb locus and another 100 kb locus; the intensity of each pixel represents the normalised number of contacts between a pair of loci (scale: 0-162). Interactions identified in wild-type cells are shown above the diagonal and interactions in *Ctcf* hemizygous cells are shown below the diagonal. No large-scale differences are identified between the two genotypes. Genes and TADs are shown; *Ctcf* is located at chr8:105,636,568-105,682,922.

3.2.7 Altered gene expression patterns are recapitulated in mouse and human tumours

In order to assess the relevance of our findings to the process of tumourigenesis *in vivo*, we asked whether CTCF-dependent cancer pathways were also activated in the transcriptomes of primary mouse and human tumours. Notably, *CTCF* is detected as a mutational driver in uterine and breast carcinomas, in which most (68%) variants are truncating mutations (Gonzalez-Perez et al., 2013; Rubio-Perez et al., 2015); therefore, our *Ctcf* hemizygous model is a good parallel to this human molecular phenotype.

First, we selected twenty-five primary liver tumours that spontaneously occurred during ageing of C3H mice, which were generated in our previous study (Connor et al., 2018). We then analysed their total RNA transcriptome, with a set of normal liver controls. We compared the set of differentially expressed genes in these mouse liver tumours with the genes perturbed by *Ctcf* hemizygosity in MEFs, and found that nearly half (47.6%) of the latter were also differentially expressed in the tumours. The majority (60%) showed concordant fold-changes (**Figure 3.10**), indicating that a large proportion of the up- and down-regulated genes in the *Ctcf*^{+/-} cells were also up- and down-regulated, respectively, in the mouse liver tumours. Notably, these concordantly altered genes retained strong enrichment for cancer-related functional terms and pathways.

We next asked whether the molecular pathways perturbed by *Ctcf* hemizygosity in mouse embryonic fibroblasts were similarly perturbed in human tumours. We identified 104 uterine and 19 breast human tumour samples from TCGA with deleterious (missense, frameshift or stop gained) mutations in at least one allele of *CTCF*. To compare the gene expression profiles across species, we restricted our analyses to those genes that are one-to-one orthologs. For both the uterine and breast cancer datasets, we observed a large overlap (~75%) between the set of differentially expressed genes in *Ctcf* hemizygous MEFs and those altered in human tumours. From these, around 65% showed concordant changes across all datasets (**Figure 3.10**), supporting a common signature of transcriptional alterations upon the loss of one functional copy of *CTCF*.

In sum, our data indicate that a small reduction in the concentration of CTCF can significantly perturb the expression of hundreds of transcripts required for normal cellular homeostasis, as evidenced by their dysregulation in a diversity of mouse and human tumours.

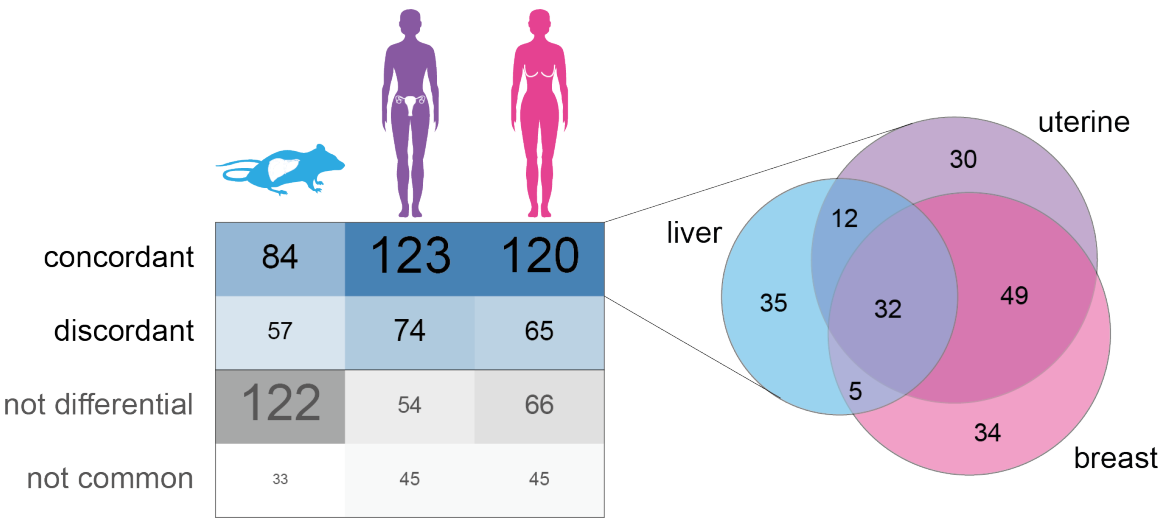


Fig. 3.10 Concordant gene alterations in diverse murine and human tumours. The comparison of the set of CTCF-dependent genes and those differentially expressed in mouse liver tumours or in human uterine and breast tumours revealed a large overlap. The majority of these changed in the same direction (concordant) in the *Ctcf* hemizygous MEFs and the tumour samples. Additionally, the set of genes unique to MEFs is indicated (not differential) and those that were either not expressed or did not have a one-to-one ortholog in the human genome (not common). The concordant gene changes across these diverse tumours are highly overlapping, as seen in the Venn diagram.

3.3 Discussion

Complete removal of CTCF has catastrophic effects as a result of massive dysregulation of the 3D genome (Lee et al., 2017; Liu et al., 2016; Schmidt et al., 2012; Zuin et al., 2014). Here we used *Ctcf* hemizyosity as a model system to compare how transcription and genome organisation in otherwise identical cells adapt to differing concentrations of CTCF. This model closely approximates the normal physiological variation of CTCF levels across tissues without the confounding effects that arise from cell-specific *trans* environments.

These data strongly suggest that mammalian cells can compensate for fluctuations in intra-cellular CTCF concentration from DNA to RNA to protein. In MEFs, 50% removal of *Ctcf* causes a 37% reduction in mRNA expression, leading to a 27% reduction in CTCF protein, and resulting in only a 2% difference in genomic occupancy. The homeostatic and functional buffering observed in our model system offers a clear explanation for how tissues that have highly variable levels of CTCF expression (Mele et al., 2015; Uhlen et al., 2015; Yu et al., 2015) nevertheless preserve CTCF genomic occupancy levels.

Our data further indicate that sub-megabase scale chromatin structures are also robust to variation in the amount of CTCF available in the nucleus. Recent studies have shown that CTCF is dispensable for the establishment of the A and B compartments but necessary for the proper insulation of TADs and the integrity of looping interactions (Nora et al., 2017). We did not observe any changes in the structure or insulation of TADs in the *Ctcf* hemizygous cells (data not shown), consistent with the high conservation of TADs observed across tissues.

The controlled reduction in CTCF expression in hemizygous cells revealed, however, reproducible changes to the nuclear environment, thus providing insights into its inherent functions. Almost a thousand loci were directly bound by CTCF that showed reproducible quantitative changes in their genomic occupancy. These were accompanied by alterations in the transcription of several hundred genes, which in turn affected the corresponding protein abundances. Since the promoters of these genes were differentially deployed between the two genotypes, these transcriptional changes arose from alterations in nuclear homeostasis not differences in transcript stability in the cytoplasm. Near these dysregulated genes there was an excess of unstable fine-scale chromatin interactions and enhancers showing altered activity connected via loops. Therefore, in contrast to high-order chromatin structures, which are indifferent to fluctuations in CTCF concentration, fine-scale genome organisation

is more sensitive, and these alterations impact the regulatory landscape leading to a perturbed functional state.

The transcriptional alterations seen here may be the result of increased variability in the expression of these genes. Indeed, the loss of promoter-enhancer interactions due to CTCF knockdown or the deletion of its binding sites can result in increased cell-to-cell variability in gene expression (Ren et al., 2017). The use of recently developed single-cell sequencing technologies would offer the opportunity to address chromatin accessibility, DNA methylation, and transcriptional variability between cells of a particular cell population (Clark et al., 2018). A reduction in CTCF results in highly reproducible changes in population of cells' epigenomes and transcriptomes, either due to increased cell-to-cell variability or gene-specific changes occurring uniformly across cells. These gene expression changes were found disproportionately in cancer-related pathways and, consistently, a large proportion of these genes are dysregulated in the transcriptomes of mouse and human tumours from diverse origins.

In conclusion, our data support the hypothesis that, although mammalian cells are tolerant to a reduced concentration of CTCF, there is specific dysregulation of oncogenic pathways that may confer an increased predisposition to cancer.

Chapter 4

Genetic and chemical models of hepatocarcinogenesis

4.1 Introduction

***CTCF* as a tumour suppressor gene**

CTCF was first proposed as a tumour suppressor gene when 16q22.1 was found to be frequently deleted in sporadic breast and prostate cancers (Filippova et al., 1998). More recently, human cancer sequencing studies have shown that tumours driven by *CTCF* have loss-of-function mutations, either missense coding alterations or transcript truncation, in keeping with the action of tumour suppressor genes (Rubio-Perez et al., 2015). Mouse models of *Ctcf* hemizyosity show that tissues with hemizygous loss of *CTCF* exhibit increased genome-wide variability in CpG methylation. Furthermore, while *Ctcf* hemizygous mice develop normally, they demonstrate a multi-lineage predisposition to cancer (Kemp et al., 2014), thus establishing *Ctcf* as a prominent tumour suppressor gene and suggesting that *CTCF*-mediated epigenetic stability acts as a major barrier to neoplastic progression. In addition to *CTCF* being implicated in tumourigenesis, *CTCF* binding sites are mutated many human cancer types (Kaiser et al., 2016; Katainen et al., 2015), including in liver cancer (Fujimoto et al., 2016; Schulze et al., 2015).

Hepatocellular carcinoma

Liver cancer is a heterogeneous group of diseases, including hepatocellular carcinoma, cholangiocarcinoma, hepatoblastoma, and hemangiosarcoma. Hepatocellular

carcinoma accounts for 70-80% of all primary liver cancers and is the second most common cause of cancer death worldwide (Torre et al., 2015). HCC develops in the context of chronic liver disease in 80% of cases. Chronic liver disease most commonly arises from chronic inflammation due to hepatitis B or hepatitis C virus infections, dietary exposure to aflatoxin B, alcoholic liver disease (ALD), or non-alcoholic fatty liver disease (NAFLD), which causes fibrosis and subsequent cirrhosis: liver scarring in the presence of regenerative hepatocytes. Such chronic hepatocyte injury leads to genetic damage that underpins HCC (Hardy and Mann, 2016). The incidence of HCC is increasing globally, most rapidly in developed countries, largely due to the rising prevalence of NAFLD (Koh et al., 2016).

The aetiology of liver cancer is diverse, which is reflected in the molecular heterogeneity of the disease and is thought to contribute to the highly variable prognosis of patients with HCC. High-throughput sequencing of hundreds of human liver tumours has identified several different oncogenic pathways and a wide range of putative driver mutations underlying hepatocarcinogenesis (Ally et al., 2017; Boyault et al., 2007; Fujimoto et al., 2016; Letouzé et al., 2017; Schulze et al., 2016; Zucman-Rossi et al., 2015). Investigating the mechanisms by which HCC progresses from a chronic fibroinflammatory or cirrhotic disease state with macroregenerative nodules (MRNs) to invasive HCC would allow us to (i) develop a more meaningful molecular classification of HCC, in turn offering the promise of predictive and prognostic biomarkers and potential therapeutic targets (personalised or precision medicine); or (ii) identify targets to prevent progression of liver disease from a benign to malignant state.

Many cancers, including HCC, carry a high burden of somatic mutations due to the action of known genotoxic insults. Identifying *bona fide* cancer drivers in the context of a dense background mutational load is challenging and requires the combination of tractable experimental models (Bakiri and Wagner, 2013) and the application of advanced computational biology approaches (Martincorena et al., 2017).

Chemical carcinogen model of liver cancer

We recently characterised a commonly-used carcinogen-driven model of liver cancer, in which tumours are induced using diethylnitrosamine in C3H mice. For the first time, we performed exome sequencing of these tumours and spontaneously arising tumours and found that: (i) DEN-induced neoplasms arising in the same mouse evolve independently; (ii) the tumours have a high, uniform number of somatic

single nucleotide variants (SNVs); (iii) DEN exposure creates a distinct mutational signature; (iv) *Hras* is the predominant, although not obligatory, oncogenic driver of hepatocellular tumours in C3H mice (Connor et al., 2018).

Although genetically-engineered mouse models of liver cancer mimic the macroscopic and microscopic features of HCC, they often lack the molecular and genetic complexities of their human counterparts. This means that although GEM models are invaluable for testing potential mechanism-based therapies, the choice of model must be carefully evaluated (Bakiri and Wagner, 2013). Such studies have been performed in other organ systems and compared the genomic landscapes of genetic and chemical mouse models of *Kras*-driven lung cancer (Westcott et al., 2014) and skin squamous cell carcinoma (Nassar et al., 2015). Human lung and skin cancer are usually driven by mutagens in tobacco or UV light, respectively, and these studies revealed that the carcinogen-induced mouse models better reflected the genetic complexity of human neoplasms.

Thus, to delineate the consequences of *Ctcf* hemizyosity *in vivo*, we chose to use our well-established DEN chemical carcinogenesis model. Since DEN-induced tumours harbour tens of thousands of mutations and CTCF binds to tens of thousands of DNA sites across the genome, combining these systems allows us to explore how the subtle differences in CTCF binding influence mutagenesis.

4.1.1 Project aim

The aim of this project was to use genetic and chemical carcinogenesis models in combination to test *in vivo* our finding that *Ctcf* hemizyosity dysregulates cancer pathways *in vitro*. This involved (i) comprehensive characterisation of *Ctcf* hemizygous mice, (ii) prolonged ageing to develop spontaneous tumours, and (iii) liver tumour induction with a chemical carcinogen. We hypothesised that DEN-induced tumours would be more frequent on a background of *Ctcf* hemizyosity.

4.2 Results

4.2.1 Characterisation of *Ctcf* hemizygous mice

Using the same *Ctcf* hemizygous mice as in Chapter 3 (**Figure 3.1**), we sought to characterise the *Ctcf*^{+/-} colony in terms of: (i) transmission of the altered *Ctcf* allele, and (ii) overt macroscopic and microscopic phenotypes, including the reported increased incidence of spontaneous tumours (Kemp et al., 2014).

Reduced transmission of the targeted *Ctcf* allele

Female C57BL/6J mice (i.e., with two intact copies of *Ctcf*) were bred with *Ctcf*^{+/-} males, also with a C57BL/6J genetic background. Based on this breeding strategy, Mendelian inheritance predicts equal numbers of wild-type and *Ctcf* hemizygous offspring. However, in contrast to our observation in the E13.5 embryos (**Table 3.1**), interrogating the inheritance of the modified *Ctcf* allele in live offspring revealed that only 30.6% of offspring in our colony were hemizygous for *Ctcf* (two-tailed t-test, $p = 3.07 \times 10^{-5}$; **Table 4.1**). This reduced transmission rate is consistent with the genotypic ratios observed in previous studies of *Ctcf* deletion and suggests that CTCF is required in a dose-dependent manner in development (Heath et al., 2008).

Table 4.1 Genotyping of *Ctcf*^{+/-} mice, listed according to maternal mouse ID.

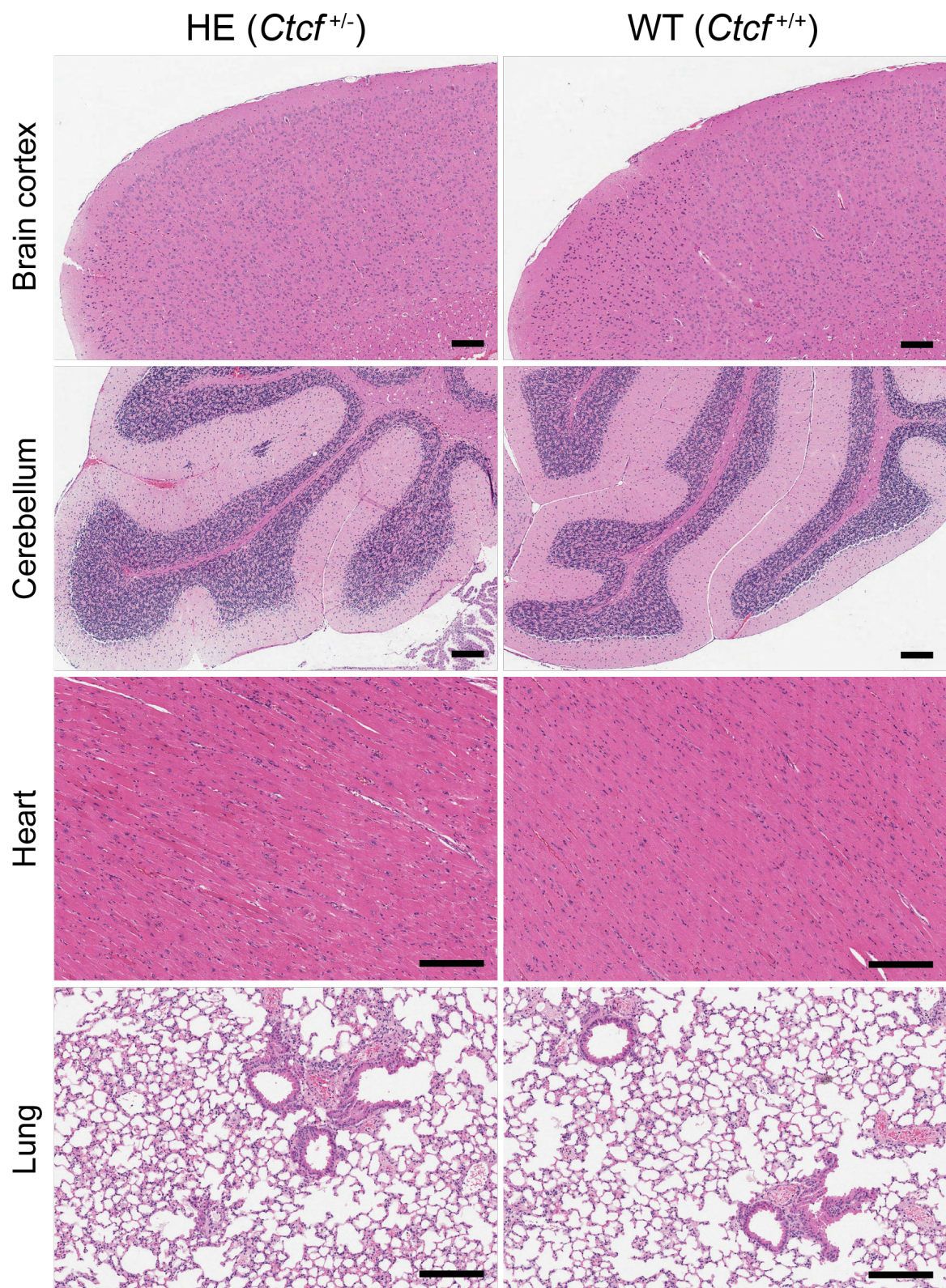
Mouse ID	Total no. offspring	WT (<i>Ctcf</i> ^{+/+})	HE (<i>Ctcf</i> ^{+/-})	% HE offspring
98683	7	6	1	14.29
97925	7	3	4	57.14
98684	5	3	2	40.00
AN14CUK035892	4	3	1	25.00
AN15CUK011173	5	5	0	0.00
AN15CUK011172	16	14	2	12.50
AN15CUK018388	5	3	2	40.00
AN15CUK025737	4	2	2	50.00
AN15CUK024738	4	3	1	25.00
AN15CUK024737	5	4	1	20.00
AN15CUK027700	9	7	2	22.22
AN15CUK027702	25	14	11	44.00
AN15CUK027701	13	8	5	38.46
AN16CUK002537	12	10	2	16.67
AN16CUK002539	8	6	2	25.00
AN16CUK002476	17	12	5	29.41
AN16CUK002538	18	16	2	11.11
AN16CUK002477	10	5	5	50.00
AN16CUK002535	7	6	1	14.29
AN16CUK015088	14	9	5	35.71
AN16CUK015087	19	16	3	15.79
AN16CUK015089	10	8	2	20.00
AN16CUK015086	27	14	13	48.15
AN16CUK015090	5	4	1	20.00
AN16CUK026234	8	6	2	25.00
AN17CUK002853	6	2	4	66.67
AN17CUK002854	5	2	3	60.00
Total	275	191	84	30.61

Histological characterisation of *Ctcf* hemizygous mice

Given the unexpectedly low number of live *Ctcf* hemizygous pups based on Mendelian principles and the suggestion that CTCF is required in a dose-dependent manner, we questioned whether there were any developmental differences between the genotypes. Both genotypes grew normally and were macroscopically indistinguishable. Although the mice had no overt phenotype, we sought to identify microscopic differences between *Ctcf* hemizygous mice and their wild-type littermates. At the time of necropsy, all major organs were dissected and processed for histological assessment.

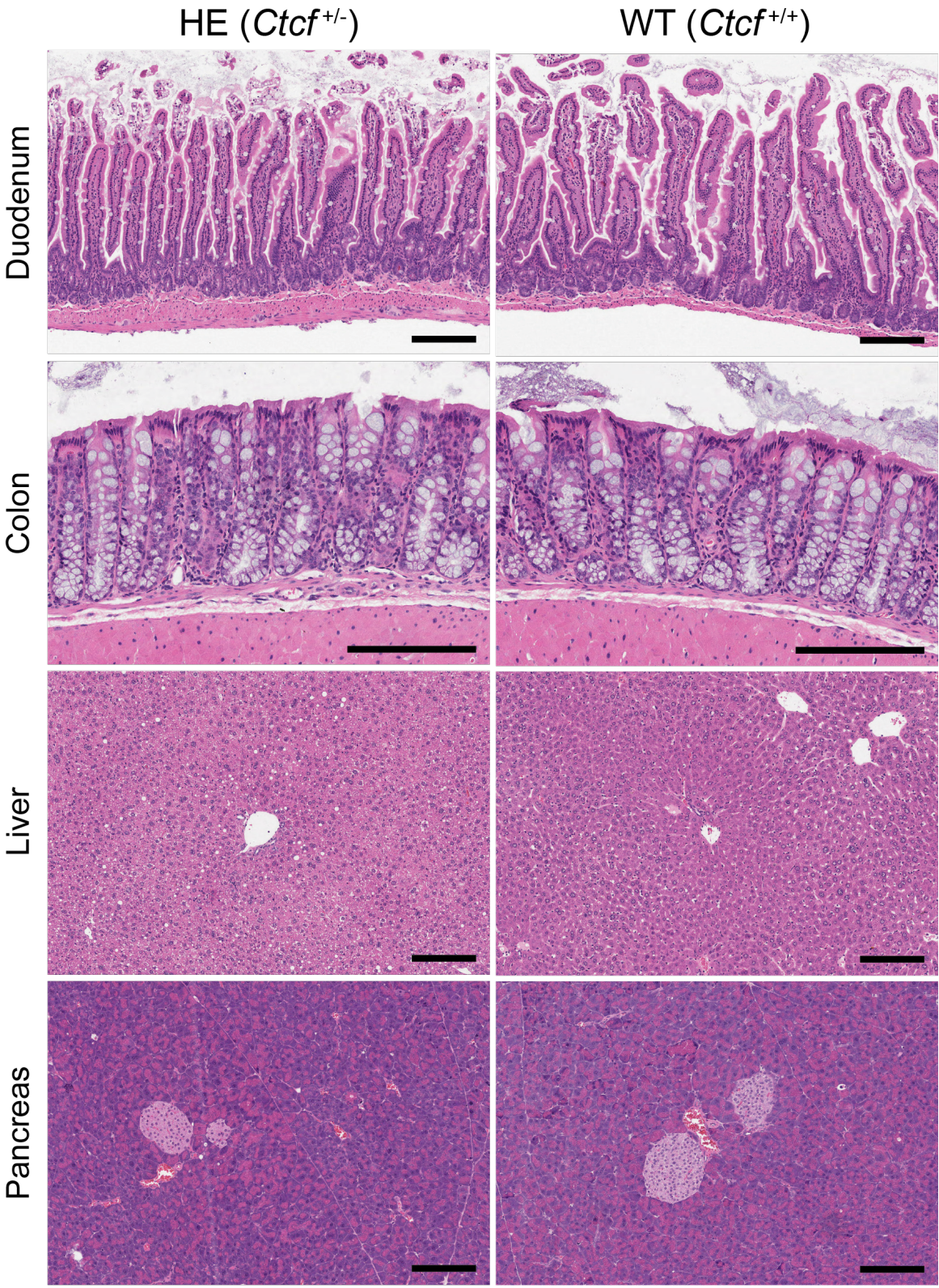
Tissue sections of heart, lung, brain, liver, gallbladder, pancreas, luminal gastrointestinal tract (oesophagus, forestomach, glandular stomach, duodenum, jejunum, ileum, caecum, colon, rectum), kidney, adrenal gland, bladder, reproductive organs, spleen, bone marrow, lymph nodes, and skeletal muscle were examined from all mice in the cohort (**Figure 4.1a-c**). Histological examination revealed normal adult development of all tissues and no morphological differences between genotypes.

In conclusion, and in keeping with Kemp et al. (2014), we did not identify any overt post-natal developmental defects in *Ctcf* hemizygous mice: pups thrived, weaned successfully, and matured normally.

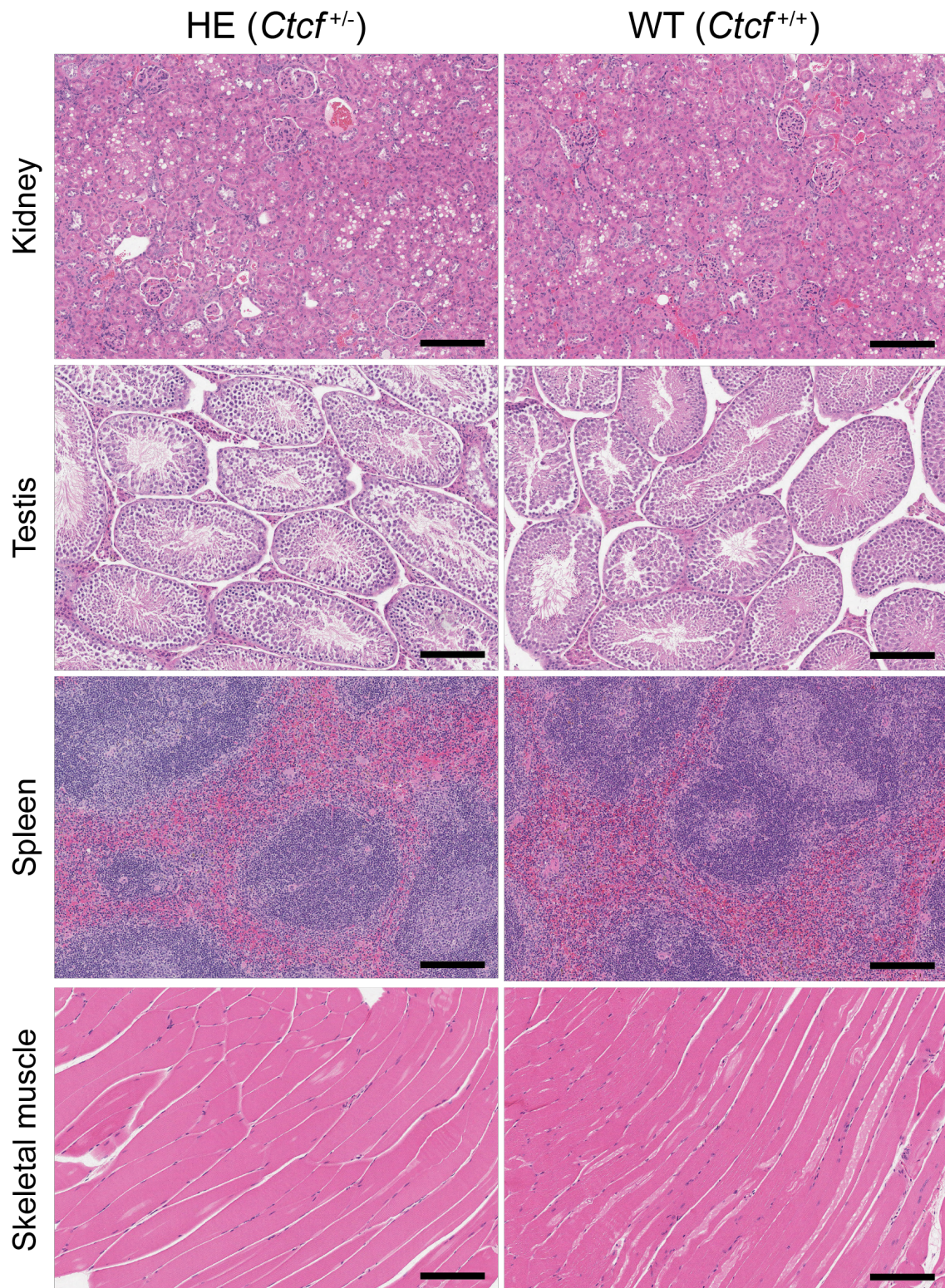


(a) Photomicrographs of tissues from *Ctcf* hemizygous mice: brain cortex, cerebellum, heart, and lung

Fig. 4.1



(b) Photomicrographs of tissues from *Ctcf* hemizygous mice: duodenum, colon, liver, and pancreas



(c) Photomicrographs of tissues from *Ctcf* hemizygous mice: kidney, testis, spleen, and skeletal muscle

Fig. 4.1 Histological analysis of tissues from *Ctcf* hemizygous and wild-type mice. Representative photomicrographs showing haematoxylin and eosin (H&E) stained tissue sections from *Ctcf* hemizygous mice and wild-type littermates. Normal morphological features are present in the (a) brain frontal cortex; cerebellum; left ventricle of the heart; lung; (b) throughout the gastrointestinal tract: including duodenum, colon, liver, and pancreas; and (c) renal cortex, testis, spleen, and skeletal muscle. All scale bars = 200 μ m. Original magnification x100.

Spontaneous tumourigenesis in *Ctcf* hemizygous mice

Kemp et al. (2014) reported that, with prolonged ageing, *Ctcf*^{+/-} mice were markedly predisposed to spontaneous tumour development in a broad range of tissues including carcinomas, sarcomas, and haematological malignancies. We therefore attempted to reproduce this tumourigenic phenotype by ageing a cohort of *Ctcf* hemizygous and wild-type mice up to 20 months old. Unexpectedly, necropsy revealed that only six out of 36 mice developed spontaneous tumours. Five of the six mice with neoplasms were hemizygous for *Ctcf* and one was wild-type (two-tailed Fisher's exact test, $p = 0.64$; **Table 4.2**).

At the time of necropsy, all macroscopically identified tumours were bisected: half of the tissue was flash frozen for DNA/RNA extraction, and the remaining tissue was processed for histopathological examination (**Figure 4.2**).

Table 4.2 Prevalence of spontaneous tumours in aged *Ctcf* hemizygous mice

	Number of mice	
	HE (<i>Ctcf</i> ^{+/-})	WT (<i>Ctcf</i> ^{+/+})
Tumours identified	5	1
Tumours not identified	19	11
Total	24	12

Histopathological characterisation of hepatocellular neoplasms

Tumours were classified according to the International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice (INHAND) guidelines (Thoolen et al., 2010). Four mice in our cohort developed liver tumours: one mouse had three tumours and the other three mice developed solitary liver tumours (**Figure 4.2**). A further two mice had solitary lung tumours.

Examination of H&E-stained sections revealed that all liver neoplasms had hepatocellular morphology with characteristic histological features indistinguishable from human liver tumours. While none of these tumours had any morphological features suggestive of cholangiocarcinoma (CC), it is possible that they could be of intermediate HCC-CC subtype. In a human diagnostic setting, this distinction can be assisted by immunohistopathology (Maximin et al., 2014): biliary cell stains include mucin, CK7, and CK19, while hepatocellular stains comprise polyclonal CEA, Hep Par 1, CD10, and glypican-3 (Maximin et al., 2014). In our mouse liver tumours, CK19 staining was negative; however, the use of additional IHC was compromised since many of the relevant antibodies are raised in mice.

Most tumours had relatively monotonous morphology with high tumour cell percentages (>85%). Dysplastic nodules (DNs) demonstrated an expansile growth pattern causing compression of adjacent hepatic parenchyma, loss of normal lobular architecture, nuclear atypia, and increased mitotic activity. HCCs showed a more abnormal spectrum of changes including thickened trabeculae, pseudoglandular structures, more marked cellular atypia, increased nuclear to cytoplasmic ratios, higher proliferative indices, and an infiltrative growth pattern. Some HCCs also had areas of haemorrhage, cystic degeneration, and/or necrosis.

In order to confidently distinguish pre-invasive DN from invasive HCCs, additional histochemical stains and immunohistochemistry (IHC) were performed on sections of liver lesions: reticulin staining highlighted the abnormal architecture of DN and was completely lost in HCCs; CD31 stained blood vessels including the endothelial lining of sinusoid-like tumour vessels (Cui et al., 1996; Sugino et al., 2008); and Ki67, which is expressed during mitosis and all other active phases of the cell cycle (G1, S, G2), is a commonly used proliferation marker (Scholzen and Gerdes, 2000). As expected, normal adult liver was mostly quiescent and all tumours showed an increased mitotic index, most markedly in HCCs. Finally, CD45 staining was used to assess the presence of tumour-infiltrating lymphocytes (TILs). This clearly identified Kupffer cells and occasional circulating lymphocytes in similar numbers in normal and neoplastic tissue. Although CD45 is a pan-leukocyte marker and not specific

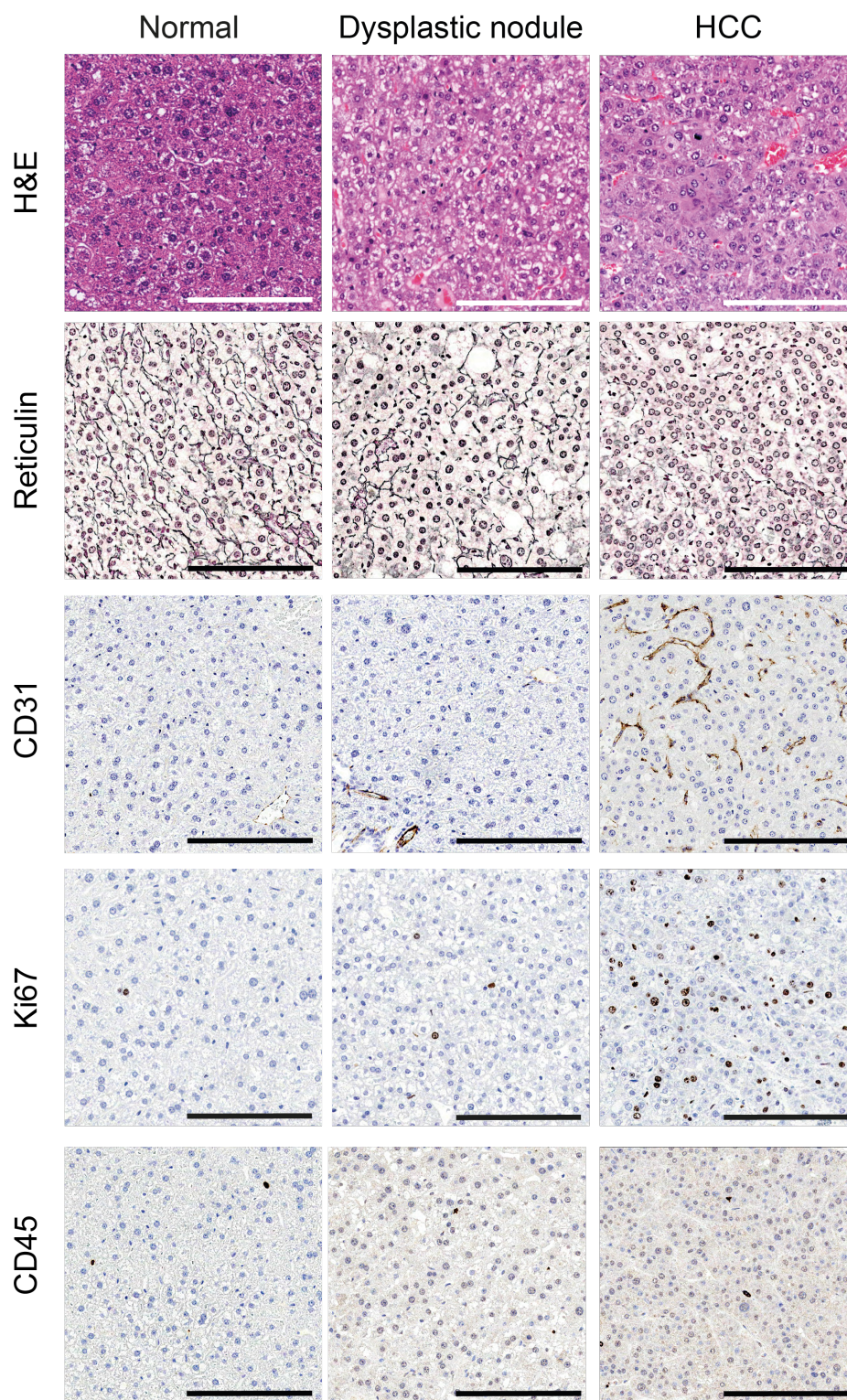


Fig. 4.2 Histological characterisation of hepatocellular neoplasms arising in *Ctcf*^{+/-} mice. Representative photomicrographs of serial sections of normal and neoplastic liver tissue. H&E staining demonstrates tissue morphology, reticulin staining is used to assess architecture, IHC with antibodies targeting CD31 demonstrates abnormal vasculature, Ki67 identifies mitotic cells, and CD45 identifies lymphocytes. All scale bars = 200 μ m. Original magnification x200.

to TILs, the negative result allowed us to confidently conclude that TILs are not a feature of these tumours.

Finally, a comparison of tumours arising in *Ctcf* hemizygous and wild-type mice revealed identical morphological features. The small number of tumours did not provide the statistical power to compare differences in the number of DNAs and HCCs arising in each genotype.

4.2.2 Hepatocyte-specific *Ctcf* knockdown

The hypothesis that *Ctcf* haploinsufficiency predisposes to carcinogenesis is based on our *in vitro* data showing dysregulation of cancer pathways (presented in Chapter 3) and published evidence of an increased incidence of tumours in *Ctcf*^{+/-} mice (Kemp et al., 2014). In order to further test this hypothesis *in vivo*, we generated a hepatocyte-specific *Ctcf* haploinsufficient mouse model and chemically induced liver tumours in these animals.

In cell type-specific models, haploinsufficiency is present in a particular tissue or organ so that it can be explored in isolation, negating the systemic immunological or metabolic effects present in germline hemizygous models that might otherwise confound the findings. We chose the liver-specific model for several reasons: (i) the cell population is more homogeneous than other tissues (~85% hepatocytes, with fewer fibroblasts, endothelial, and inflammatory cells); (ii) it is well characterised by previous studies in our laboratory (Odom et al., 2007; Schmidt et al., 2010a,b; Schwalie et al., 2013; Villar et al., 2014); (iii) we have significant experience with a liver tumour-induction protocol (Connor et al., 2018); (iv) the liver is easily identified at necropsy; and (v) a single mouse liver provides sufficient tissue for several assays. Such conditional knock-out mice rely on high expression levels of a cell-type specific protein. Since albumin is highly and uniquely expressed in hepatocytes, we exploited B6.Cg-*Speer6-ps1*^{Tg(*Alb-cre*)21Mgn}/J mice (Postic et al., 1999) (hereafter referred to as: *Alb-cre*), which express Cre recombinase under the control of the mouse albumin promoter, to target *loxP*-flanked genes in the liver. *Alb-cre*-positive female mice were bred with *Ctcf*^{flox/flox} males to produce hepatocyte-specific *Ctcf* hemizygous knockout mice. Their wild-type littermates were used as controls (**Methods**).

Genotyping of ear biopsies revealed that this breeding regimen generated the expected 1:1 Mendelian ratio of offspring, with balanced numbers of pups with hemizygous deletion of *Ctcf* in hepatocytes (B6.*Ctcf*^{tm1.1Laat}/*Ctcf*⁺; *Speer6*-

Table 4.3 Genotyping of liver-specific *Ctcf* hemizygous mice. Mice are listed according to maternal mouse ID including the total numbers of female and male offspring and genotypic breakdown of the male mice used for experiments

Mouse ID	Female mice	Male mice			
	Total no.	Total no.	HE;WT	HE;HE	HE;HE (%)
AN15CUK011743	3	6	2	4	66.67
AN15CUK011748	2	5	2	3	60.00
AN15CUK016208	0	7	5	2	28.57
AN15CUK016492	1	9	4	5	55.56
AN15CUK016501	12	3	2	1	33.33
AN15CUK016503	5	3	1	2	66.67
AN15CUK020057	0	3	2	1	33.33
AN15CUK020059	6	4	4	0	0.00
AN15CUK020240	4	3	0	3	100.00
AN15CUK020241	5	0	0	0	-
AN15CUK020714	4	3	2	1	33.33
AN15CUK020863	9	1	1	0	0.00
AN15CUK021703	3	0	0	0	-
AN15CUK024953	4	5	2	3	60.00
AN15CUK025142	1	2	0	2	100.00
AN15CUK026801	3	5	4	1	20.00
AN15CUK027412	2	5	3	2	40.00
AN15CUK028660	5	4	1	3	75.00
AN15CUK028661	4	3	1	2	66.67
AN15CUK029396	3	7	2	5	71.43
AN15CUK029403	5	4	2	2	50.00
AN15CUK029705	4	5	2	3	60.00
AN15CUK029919	3	6	2	4	66.67
AN15CUK030711	3	5	2	3	60.00
AN15CUK030712	1	0	0	0	-
AN15CUK030713	3	2	1	1	50.00
AN15CUK030714	4	3	0	3	100.00
AN15CUK032186	6	3	2	1	33.33
AN15CUK033231	5	3	0	3	100.00
AN15CUK033233	2	2	1	1	50.00
AN15CUK033939	5	2	1	1	50.00
AN15CUK033940	4	5	1	4	80.00
AN15CUK036396	6	3	3	0	0.00
AN15CUK037934	3	3	3	0	0.00
AN15CUK037935	5	1	1	0	0.00
AN15CUK040215	3	4	2	2	50.00
AN15CUK041041	7	0	0	0	-
AN15CUK045623	0	6	4	2	33.33
AN16CUK000627	3	5	3	2	40.00
AN16CUK000628	3	4	2	2	50.00
AN16CUK000972	6	3	3	0	0.00
AN16CUK001248	5	4	3	1	25.00
AN16CUK001251	4	11	6	5	45.45
AN16CUK001256	1	2	2	0	0.00
AN16CUK001257	4	7	5	2	28.57
AN16CUK001258	3	6	4	2	33.33
AN16CUK001263	6	2	1	1	50.00
AN16CUK001264	3	7	3	4	57.14
AN16CUK001265	4	4	3	1	25.00
AN16CUK002780	0	4	4	0	0.00
AN16CUK002782	0	6	4	2	33.33
AN16CUK005346	0	3	2	1	33.33
AN16CUK005348	0	5	5	0	0.00
AN16CUK009302	2	4	2	2	50.00
AN16CUK009304	2	4	3	1	25.00
Total	191	216	120	96	44.44

ps1^{Tg(Alb-cre)21Mgn}, "HE;HE") and wildtype littermates (B6.*Ctcf^{tm1.1Laat}/Ctcf⁺*, "HE;WT") (two-tailed t-test, $p = 0.104$, **Table 4.3**).

Since *Ctcf* gene expression in the liver was inferred from genotyping the presence or absence of the germline *Alb-Cre* status, it was essential to confirm recombinase efficacy in the liver. Genomic DNA from the tail and liver was analysed by qPCR, which confirmed successful recombination of the genetically altered *Ctcf* locus in liver tissue but not in the tail (**Figure 4.3A**). CTCF protein levels were quantified by western blotting, which demonstrated reduced CTCF expression in liver compared with pancreas and lung (**Figure 4.3B**). More specifically, IHC for CTCF protein in liver tissue demonstrated depleted expression in hepatocytes but not in endothelial cells, Kupffer cells, or lymphocytes (**Figure 4.3C**).

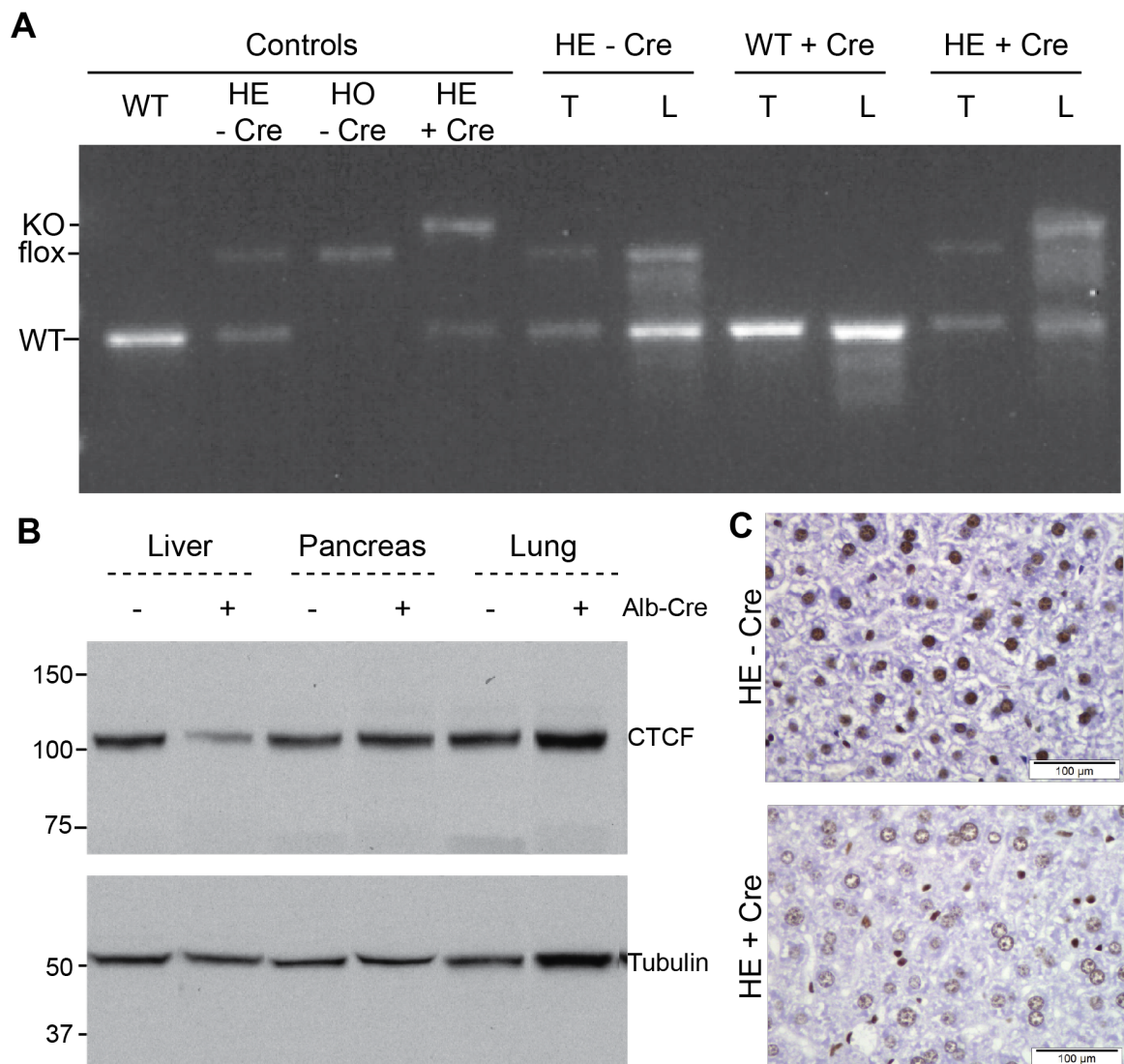


Fig. 4.3 Hepatocyte-specific *Ctcf* knockdown. Validation experiments showed that hepatocyte-specific *Ctcf* knockdown was successful. Recombination of the *Ctcf* allele from genomic DNA was demonstrated, and consequently there was reduced CTCF protein expression in hepatocytes. (A) Expected PCR product sizes for wild-type *Ctcf* (WT), *Ctcf* with loxP sites (flox), and excised *Ctcf* (KO) are shown for control samples from wild-type (WT) mice, heterozygous (HE), or homozygous (HO) for floxed *Ctcf* alleles. Recombination of the *Ctcf* allele in genomic DNA from the tail (T) and liver (L) is shown for WT and HE mice in the presence (+) or absence (-) of *Alb-cre*. (B) Western blots show tissue-specific reduction in CTCF protein expression in the liver compared with pancreas and lung. (C) Immunohistochemistry for CTCF in liver tissue from mice heterozygous for the floxed *Ctcf* allele in the absence (-) and presence (+) of *Alb-cre*, demonstrating a hepatocyte-specific reduction in CTCF expression.

4.2.3 Tumour induction using diethylnitrosamine

We recently characterised the exome-wide pattern of mutations in diethylnitrosamine (DEN)-induced tumours from C3H mice and found that: (i) DEN-induced neoplasms arising in the same mouse evolved independently; (ii) the tumours had a high, uniform number of somatic single nucleotide variants (SNVs); (iii) DEN exposure created a distinct mutational signature; and (iv) *Hras* was the predominant, although not obligatory, oncogenic driver of hepatocellular tumours (Connor et al., 2018). We have now employed a similar approach in hepatocyte-specific *Ctcf* hemizygous and wild-type mice on a C57BL/6J background and applied whole-genome sequencing (WGS) to gain insights into the genome-wide mutational landscape and to explore the impact of *Ctcf* hemizygosity in this model.

Chemically-initiated liver tumours were generated by treating 15 day-old male mice with a single intra-peritoneal dose of DEN and aged until tumours developed (**Methods**). Mice were euthanised, full necropsies were performed, livers were isolated, and tumours were dissected together with adjacent normal tissue. Half of the tumour was flash frozen (FF) for DNA/RNA extraction and sequencing, and the other half (with adjacent normal tissue) was processed in parallel for histological examination (**Figure 4.4**).

C57BL/6 mice are relatively resistant to tumourigenesis (Puccini et al., 2013), and the literature reports a broad range in tumour latency after the administration of DEN (Verna et al., 1996). For this reason, we performed an initial pilot experiment to establish the optimal time-points to collect tumours from our mouse cohort (**Table 4.4**). Based on this, we selected 36 weeks post-treatment as the primary time to collect samples: this was the earliest time at which we could confidently expect to

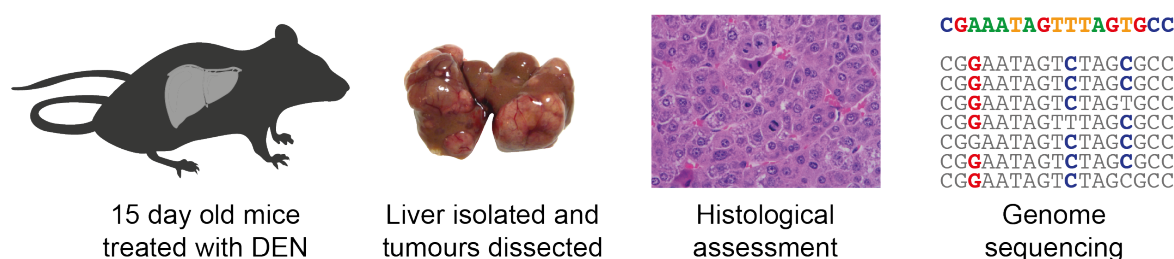


Fig. 4.4 Overview of tumour induction protocol. 15 day-old mice were treated with diethylnitrosamine (DEN) and aged until tumours developed. The liver was isolated, and tumours were dissected along with adjacent normal tissue. Tumours were bisected and processed in parallel for histological assessment and genome sequencing.

find tumours of adequate size to provide enough tissue to be bisected and processed in parallel for both histological assessment and genomic characterisation. In addition, we aged 20 mice of each genotype to 42 weeks post-treatment to see whether the phenotypic differences between the two genotypes became more pronounced with prolonged ageing.

Table 4.4 Pilot experiment to determine the optimal time point for sample collection. Two mice were euthanised at each time point to assess the number of liver tumours. *the second mouse was excluded due to an unrelated health concern.

Weeks post-DEN treatment	Wild-type		<i>Ctcf</i> hemizygous	
	No. mice	No. with tumours	No. mice	No. with tumours
19	2	0	2	0
23	2	0	2	0
27	2	2	1*	1
29	2	1	2	1
31	2	2	2	1
33	2	1	2	2
35	2	2	2	2

Mice of both genotypes developed multiple discrete tumours by 36 weeks after DEN treatment, concordant with previous studies (Heindryckx et al., 2009). Mice were euthanised and tumours collected either 36 or 42 weeks after treatment. Mouse body and organ weights were recorded at necropsy. Liver-specific *Ctcf* hemizygous mice aged for 42 weeks were heavier (mean 42.5 g) than their wild-type litter mates (mean 39.4 g, two tailed t-test, $p = 0.049$, **Figure 4.5A**). This was even more pronounced in the liver weights (4.5 g vs. 3.3 g, two tailed t-test, $p = 0.007$, **Figure 4.5B**). There was no difference in the weights of the brains, hearts, or kidneys

There was a wide range in the frequency (1 - 43 tumours per mouse) (**Figure 4.5C**) and size (1 - 23 mm diameter) of tumours isolated from mice of both genotypes (**Figure 4.5D**). At each time point, the total number of tumours (two tailed t-test at 36 weeks, $p = 0.73$ and 42 weeks, $p = 0.33$), number of tumours large enough (>2 mm) to sample ($p = 0.40$, $p = 0.15$), average tumour size ($p = 0.94$, $p = 0.30$), and tumour burden (calculated as cumulative tumour diameter, $p = 0.77$, $p = 0.07$) were comparable between the two genotypes.

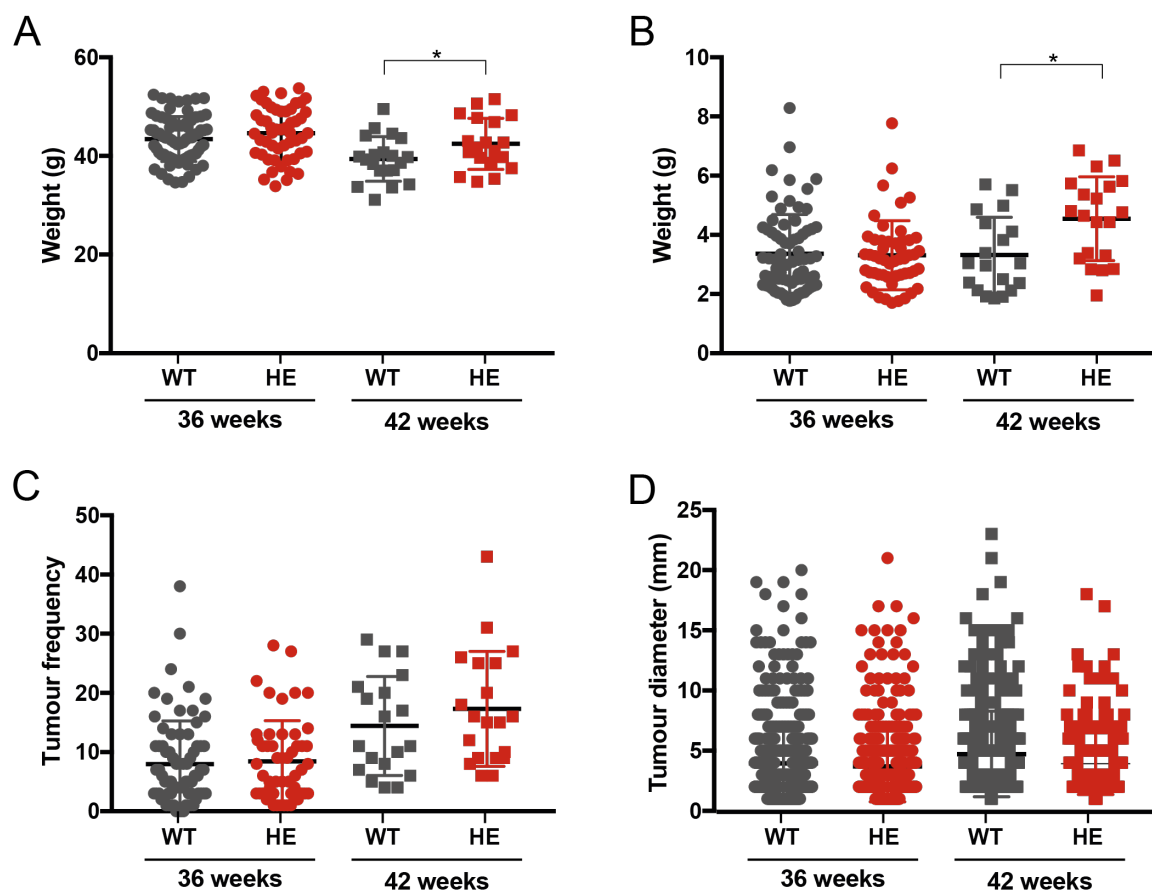


Fig. 4.5 DEN-initiated tumour characteristics. Wild-type (WT, grey) and liver-specific *Ctcf* hemizygous (HE, red) mice were treated with DEN. Liver tumours were sampled 36 weeks (round markers) or 42 weeks (square markers) after treatment. (A) Mouse body weight (g) at time of necropsy. (B) Total liver weight, including tumours, at time of necropsy. (C) Number of tumours identified macroscopically, frequency of tumours plotted per mouse. (D) Liver tumour diameter, measured at necropsy. Bars indicate mean \pm SD. Two tailed t-tests, * $p < 0.05$.

4.2.4 Pathological characterisation of DEN-induced tumours

All macroscopically identified tumours were processed for histological assessment using the same INHAND criteria described above (Thoolen et al., 2010). DEN-induced tumours all had a hepatocellular phenotype, and the full spectrum of pathological entities were present across the cohort including microscopic basophilic and eosinophilic foci of cellular alteration (phenotypically distinct hepatocytes that are potential neoplastic precursors), low- and high-grade dysplastic nodules, and well-, moderately-, and poorly-differentiated HCCs. A subset of tumours had a nodule-in-nodule appearance, supporting the hypothesis of stepwise progression from DN

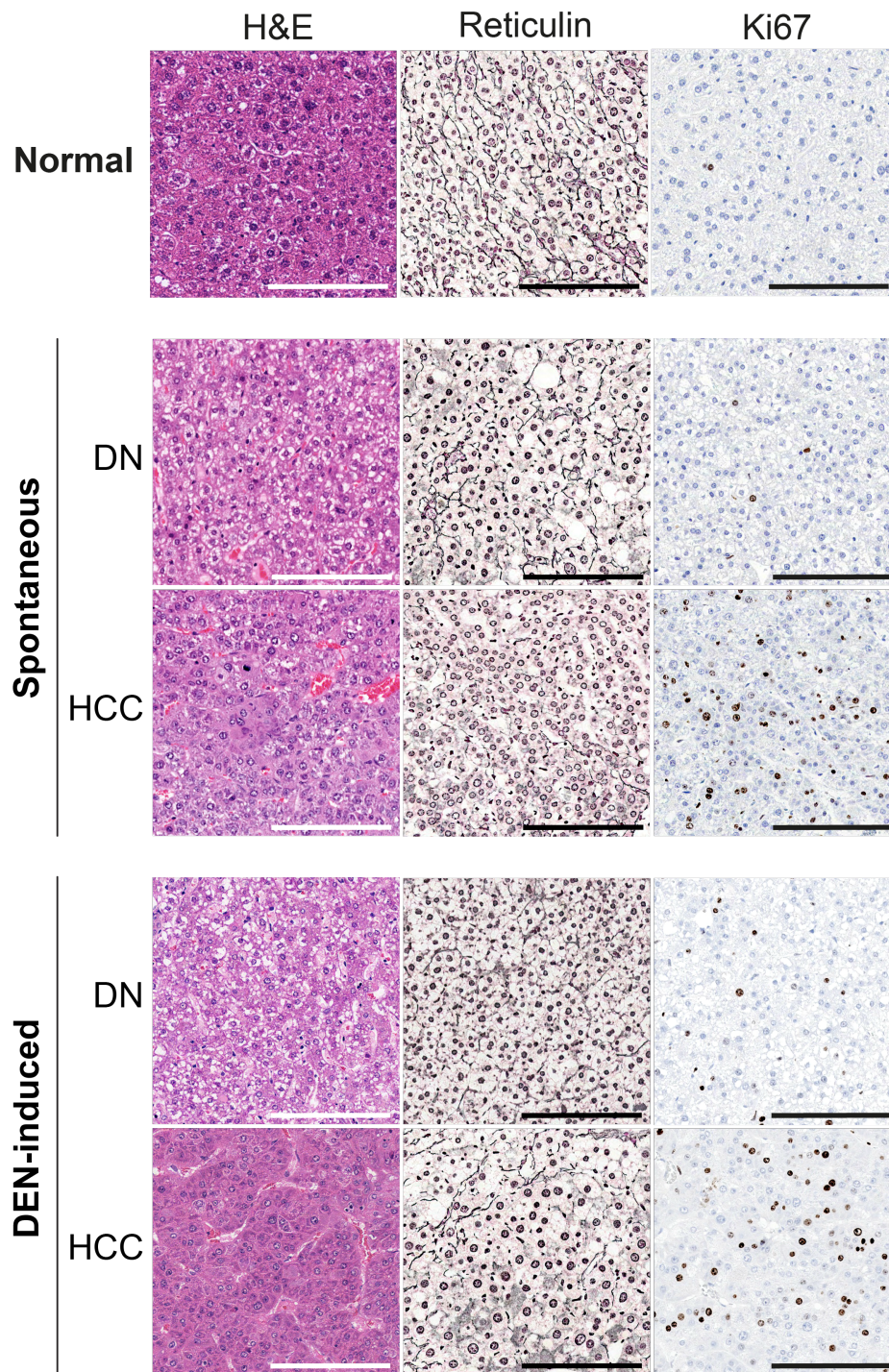


Fig. 4.6 Histology of hepatocellular neoplasms. Representative photomicrographs of serial sections of normal liver tissue and indistinguishable liver tumours arising in DEN-treated and untreated mice (DN: dysplastic nodule; HCC: hepatocellular carcinoma). H&E staining demonstrates tissue morphology; reticulin staining is used to assess architecture; and Ki67 identifies mitotic cells. All scale bars = 200 μ m. Original magnification x200. Adapted from Connor et al. (2018).

to HCC (Nam et al., 2005). Tumours arising in wild-type mice and those from *Ctcf* hemizygous mice were indistinguishable. Furthermore, DEN-induced tumours were histologically indistinguishable from spontaneous tumours (**Figure 4.6**).

HCC in humans usually occurs in the context of chronic inflammation (El-Serag and Rudolph, 2007). In contrast, there were only scattered lymphocytes present in the tumours and background liver of both genotypes in our DEN-treated cohort, with no significant immune component in the tumours (Schneider et al., 2012). A subset of tumours had prominent extramedullary haematopoiesis (EMH), which is a normal feature of liver physiology in adult C57BL/6 mice (Thoolen et al., 2010).

Macroscopically, livers from *Ctcf* hemizygous mice appeared, on average, paler than their wild-type littermates. This was also reflected in the microscopic appearance of the background liver but not the tumours (**Figure 4.7**). This is in keeping with the finding that *Ctcf* hemizygous mice had heavier, fattier livers and higher overall bodyweight.

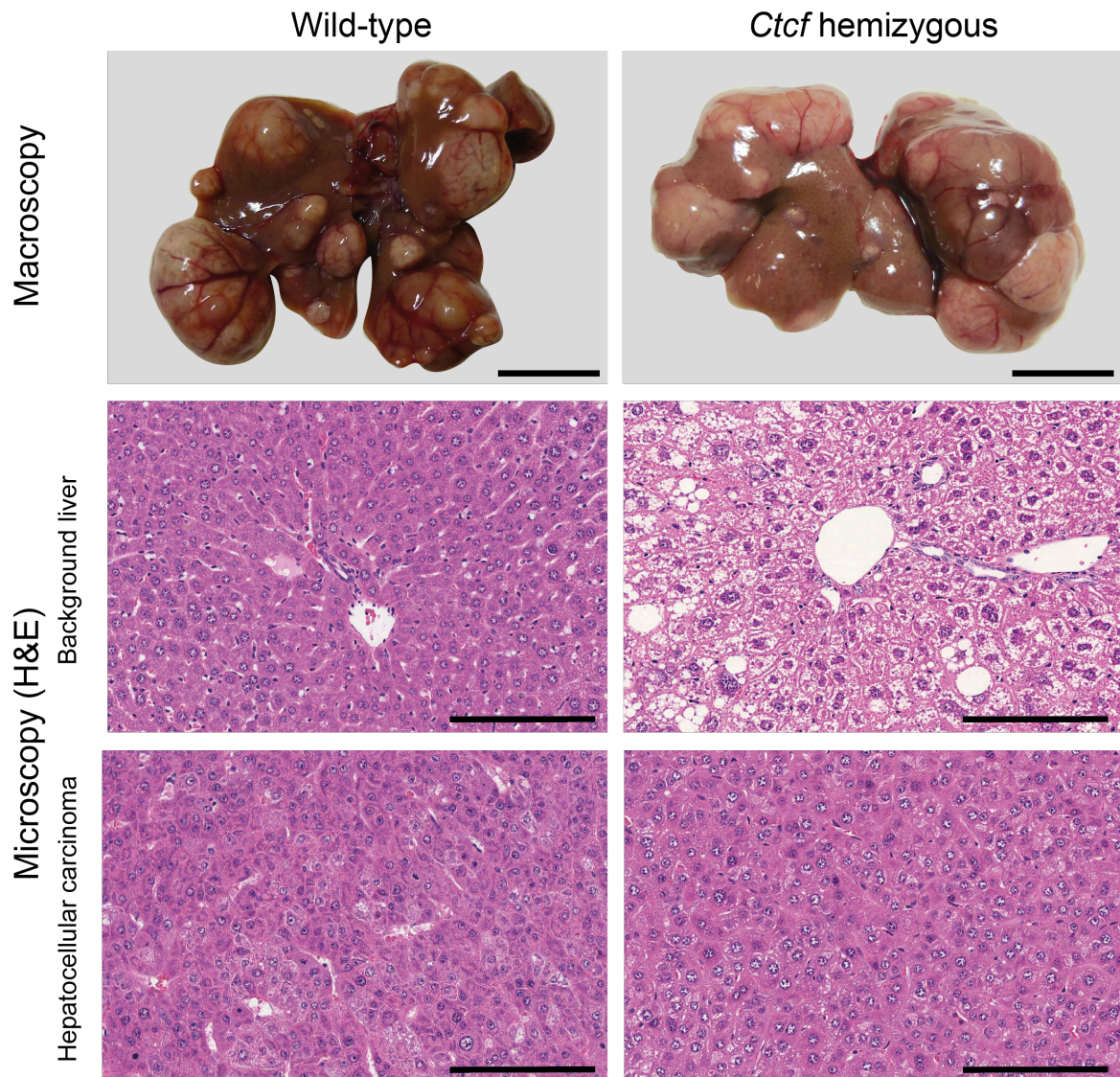


Fig. 4.7 Macroscopic and microscopic appearance of DEN-initiated liver tumours. Macroscopically, the liver parenchyma of *Ctcf* hemizygous mice was paler than their wild-type littermates and the visceral surface had a "nutmeg" appearance. Microscopic appearance (H&E) of the background liver of *Ctcf* hemizygous mice showed steatotic changes, but HCCs were indistinguishable between genotypes. Macroscopy scale bars = 10 mm. Microscopy scale bars = 200 μ m, original magnification x200.

4.2.5 Genomic characterisation *Ctcf* hemizygous liver tumours

Patients with liver cancer often present with late-stage disease and multiple liver tumours; most commonly one primary HCC with multiple intra-hepatic metastases, but synchronous multi-focal primary tumours also occur (Zeng et al., 2012). In our mouse model, most individuals also developed more than one liver tumour following DEN treatment (mean = 8 tumours (36 weeks), 16 tumours (42 weeks), **Figure 4.5C**). However, unlike their human counterparts, we recently demonstrated that these DEN-induced tumours initiate and evolve independently, since tumours arising from the same DEN-treated mouse were as genomically divergent as those isolated from separate mice (**Figure 4.8** (Connor et al., 2018)). Therefore, in this project we considered each tumour, rather than each mouse, as an independent sample.

In Chapter 3 we showed that *Ctcf* hemizygosity dysregulates cancer pathways, allowing us to hypothesise that conditional *Ctcf* knockdown in the liver would have a molecular phenotype when treated with DEN. Therefore, WGS was performed on a total of 96 samples; whenever possible, a paired normal liver sample and tail/ear from the same mouse was included in addition to the tumours. All of the spontaneously occurring liver tumours and a subset of mouse samples collected 36 weeks after treatment with DEN (**Methods**) were included. Six *Ctcf* hemizygous and six wild-type mice were selected and, for each of these mice, DNA was sequenced from 4-6 tumours (including DNs and HCCs from each mouse), background DEN-exposed liver tissue, and a tail sample. The non-tumour liver samples were included to assess the background mutational burden persisting in tissue that had been exposed to the active metabolite but had not undergone neoplastic clonal expansion. The tail samples were included to identify germline SNPs or other artefacts resulting from the compound or genetically altered mouse line. We did not expect any DEN-related genomic variants in the latter samples, since DEN is metabolised to its active form by cytochrome P450 enzymes (Kang et al., 2007) in centrilobular hepatocytes (Oinonen and Lindros, 1998), where the resulting DEN metabolites directly damage DNA by alkylating nucleobases (Verna et al., 1996).

The SNV rate was calculated for each tumour to compare the frequency of mutations in *Ctcf* hemizygous and wild-type tumours. The mutation rate (28.5/Mb) was the same in DEN-induced tumours arising in *Ctcf* hemizygous and wild-type tumours (**Figure 4.9A**). This very high mutational burden was in stark contrast to the very low frequency of SNVs in spontaneously arising tumours, in spite of indistinguishable histopathology, similar to that observed in human liver cancers (Schulze et al., 2015). Tumours of both genotypes had very few somatic indels or copy number variants

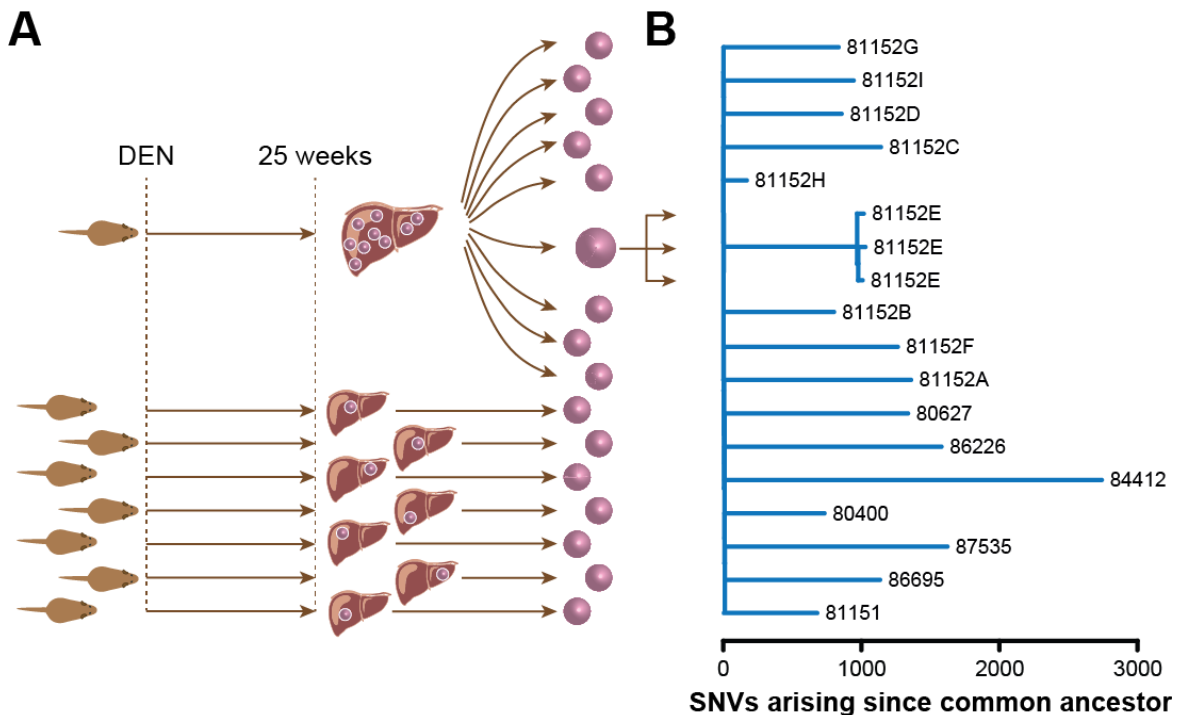


Fig. 4.8 Independent evolution of DEN-induced liver tumours is revealed by their unique SNV profiles. Connor et al. (2018) sampled liver tumours 25 weeks after DEN treatment. (A) Whole exome sequencing of nine nodules isolated from a single liver, and single nodules from seven other mice was performed. To evaluate technical noise, triplicate sequencing libraries were prepared for a single nodule. (B) A phylogenetic tree was constructed in R using the ape package (v3.5 (Paradis et al., 2004)), in which branch lengths correspond to the number of unshared SNVs (Saitou and Nei, 1987). Long branch lengths indicate no relatedness among the nodules within a single mouse, whereas three replicate libraries from the single tumour had short branches, indicating few SNV differences. Branches are labelled using mouse/tumour identification codes. Adapted from Connor et al. (2018).

(CNVs), again consistent with findings in human tumours in which the number of copy number alterations in a sample is approximately anticorrelated with the number of somatic mutations in a sample (Ciriello et al., 2013).

We also compared the mutation rate in dysplastic nodules and invasive HCCs, but again there was no significant difference between the groups. The similar mutation rate between DNs and HCCs may be because the vast majority of somatic variants in both tumour types are caused by an initial single burst of mutagenesis upon DEN exposure in the originating cell. Only a small number of additional secondary driver mutation(s) are necessary for a DN to progress to a HCC, and this acquisition

and subclonal expansion may be masked by the very high burden of pre-existing passenger SNVs in the DEN-initiated DNAs.

Notably, the high mutational burden of SNVs found in the DEN-initiated tumours was comparable to human cancer cohorts with strong mutagenic drivers such as lung cancer and melanoma, where tobacco and UV light, respectively, are known to cause a very high prevalence of somatic mutations (Alexandrov et al., 2013).

The SNV rate varied markedly across the cohorts, so we considered whether tumours arising from a single mouse might have a similar mutational burden, either due to dysregulation of a DNA repair pathway (causing a hyper-mutation effect) or due to technical error in the administration of DEN. We found that the frequency of somatic variants was very consistent between tumours arising in a single mouse, with a small proportion of tumours in each mouse accounting for the spread in the data (**Figure 4.9B**). Again, the twenty-fold difference in SNV rate between DEN-induced and spontaneously arising tumours was conspicuous.

4.2.6 Distinct mutational signatures of liver tumours

DEN is metabolised to its active form by cytochrome P450 enzymes (Kang et al., 2007) in centrilobular hepatocytes (Oinonen and Lindros, 1998), where the resulting DEN metabolites can directly damage DNA by alkylating nucleobases (Verna et al., 1996). We have shown experimentally that the promutagenic O⁶-ethyl deoxyguanosine adduct accumulates in zone 3 hepatocytes within four hours of treatment with DEN in mice (Connor et al., 2018).

DEN-initiated tumours in both *Ctcf* hemizygous and wild-type mice had distinct and reproducible mutational profiles including all possible somatic base substitutions. Transitions and transversions at A and T base pairs were especially prominent (**Figure 4.10A**), consistent with persistent alkylated thymidine lesions caused by metabolically activated DEN (Verna et al., 1996). More specifically, these transversion events occurred more frequently when the T (or A) was preceded by a C (or A) and followed by a T (or G). In contrast, there was a paucity of C:G to G:C transversions in DEN-treated samples.

Tumours arising spontaneously in untreated mice not only had far fewer SNVs but also had a different distribution of mutations that was dominated by C:T variants. The mutational pattern in spontaneously arising tumours was very heterogeneous, analogous to the portraits observed in human HCCs (**Figure 4.10A**), which show marked aetiological and molecular heterogeneity (Letouzé et al., 2017).

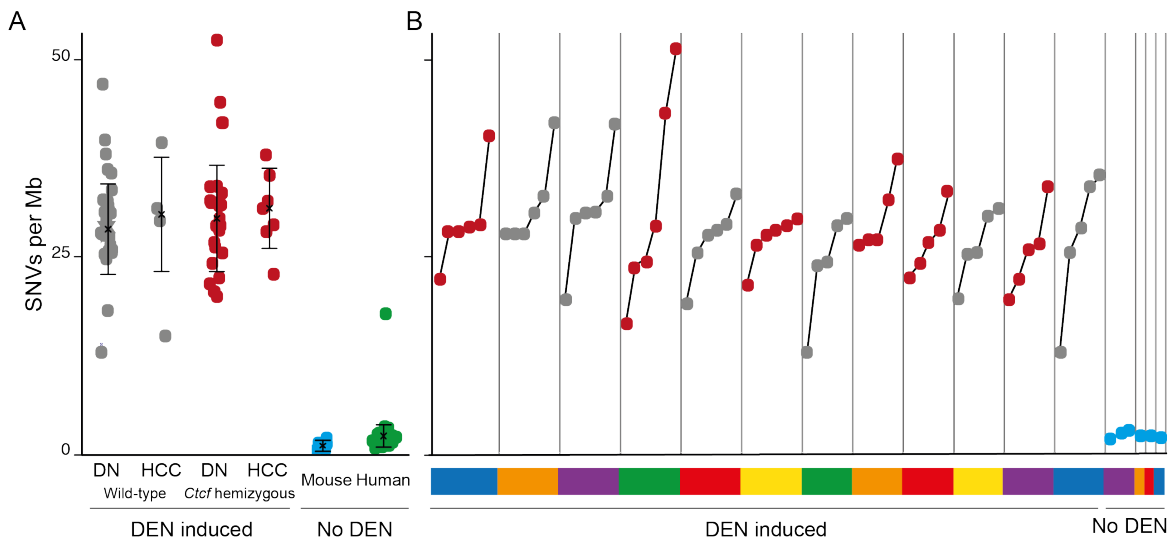


Fig. 4.9 DEN-initiated neoplasms have a high SNV burden. The single nucleotide variant (SNV) frequencies are shown for each mouse and human liver tumour cohort. These include DEN-induced dysplastic nodules (DNs, $n = 52$) and hepatocellular carcinomas (HCCs, $n = 14$) arising in *Ctcf* hemizygous (red, $n = 34$) and wild-type mice (grey, $n = 32$), and spontaneous tumours (blue, $n = 6$) arising in untreated mice. Previously reported human HCC (green, $n=50$) samples are shown for comparison (ICGC LICA-FR). Each point represents a single tumour sample. (A) The mutation rate is the same in *Ctcf* hemizygous and wild-type tumours, and there is no difference between the mutation rate between DN and HCCs. Spontaneous mouse tumours have a comparable SNV rate to human HCCs. (B) SNV frequency varies between tumours arising in the same mouse liver. Mice are ranked according to median SNV count from highest (left) to lowest (right), showing that tumours do not cluster according to genotype. Multiple tumours were sequenced from each mouse, those from the same mouse are coloured as follows. WT: 15/27262 (red), 15/31651 (orange), 15/33905 (yellow), 15/39018 (green), 15/44043 (blue), 15/44203 (purple); *Ctcf* hemizygous: 15/41619 (red), 15/44139 (orange), 15/44139 (yellow), 15/44042 (green), 15/44044 (blue), 15/44198 (purple); Spontaneous: 14/34745 (red), 14/34747 (orange), 15/28671 (green), 15/29027 (purple).

Hierarchical clustering of the tumours according to the 96 possible trinucleotide substitution contexts showed a consistent mutational pattern shared between tumours arising in *Ctcf* hemizygous and wild-type mice exposed to DEN. The phylogenetic tree demonstrated that tumours from the same mouse were no more or less similar than tumours from another individual mouse (**Figure 4.10B**). In contrast, the distribution of mutations occurring in spontaneously arising tumours is more heterogeneous and, although the sample size is small, it is clearly distinct from the mutational imprint of carcinogen-initiated neoplasms.

Given this distinct mutational pattern, we sought to define a "DEN signature" reflecting the consequence of exogenous DEN treatment and the DNA damage, repair, and replication that follows. Since our sample size was too small to derive *de novo* signatures, we reconstructed composite signatures using the profiles of known mutational signatures (Alexandrov et al., 2013; Nik-Zainal et al., 2012). The current set of thirty mutational signatures is based on an analysis of 10,952 exomes and 1,048 whole-genomes across forty distinct human cancer types (Alexandrov et al., 2015). The proportions of COSMIC mutational signatures (Forbes et al., 2017) represented in the mutational profile from each sample were calculated using the R package deconstructSigs (Rosenthal et al., 2016). Reconstruction of mutational portraits for each DEN-induced tumour from both wild-type and *Ctcf* hemizygous mice demonstrated that they were composed almost entirely of six COSMIC signatures (**Figure 4.11**). In contrast, spontaneous tumours in mice and HCCs in the human (ICGC LICA-FR) cohort were more heterogeneous.

The six COSMIC signatures that dominated the "DEN signature" were 8, 12, 21, 22, 24, and 30, half of which are found in human liver cancers: signature 12, signature 22, and signature 24. Two of these are particularly interesting, since their proposed aetiologies are clinically tractable to exogenous exposures. Signature 22 has been found in cancers with known patient exposures to aristolochic acid and is also consistent with profiles seen in experimental systems exposed to aristolochic acid, and signature 24 has been found in samples with known exposures to aflatoxin (Fujimoto et al., 2016; Letouzé et al., 2017; Schulze et al., 2015). Although signature 12 does not have a proposed aetiology, it usually only contributes a small percentage (<20%) of the mutations observed in a liver cancer sample. None of the remaining three signatures (8, 21, and 30) currently have a proposed aetiology.

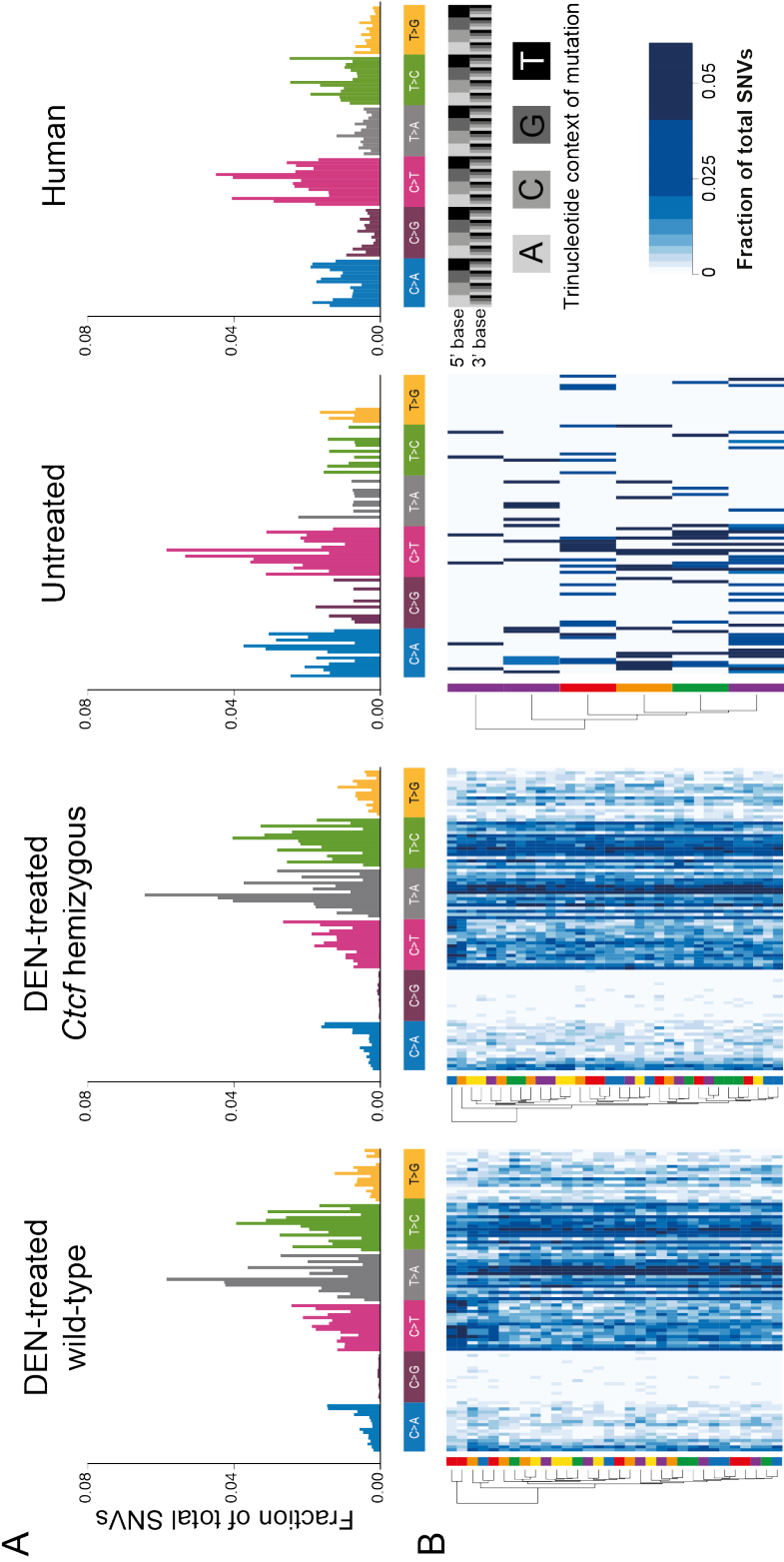


Fig. 4.10 DEN-initiated tumours have distinct and reproducible mutational profiles. (A) Frequencies of substitution mutations occurring across the cohorts of DEN-induced tumours (WT $n=32$, *Ctcf*^{+/−} $n=34$); spontaneous tumours arising in untreated mice ($n=6$); and human liver tumours (ICGC LICA-FR, $n=50$). SNVs are shown using the 96 substitution classification, which is defined by reporting the specific base substitution combined with the immediate neighbouring 5' and 3' nucleobases. (B) Heatmaps displaying the mutational profiles of each mouse tumour (rows) classified by substitution and trinucleotide context (columns). The phylogenetic tree to the left of each heatmap quantifies the clustering observed from the individual mouse sample mutational profiles. Multiple tumours were sequenced from each mouse; those from the same mouse are colour coded as per Figure 4.9.

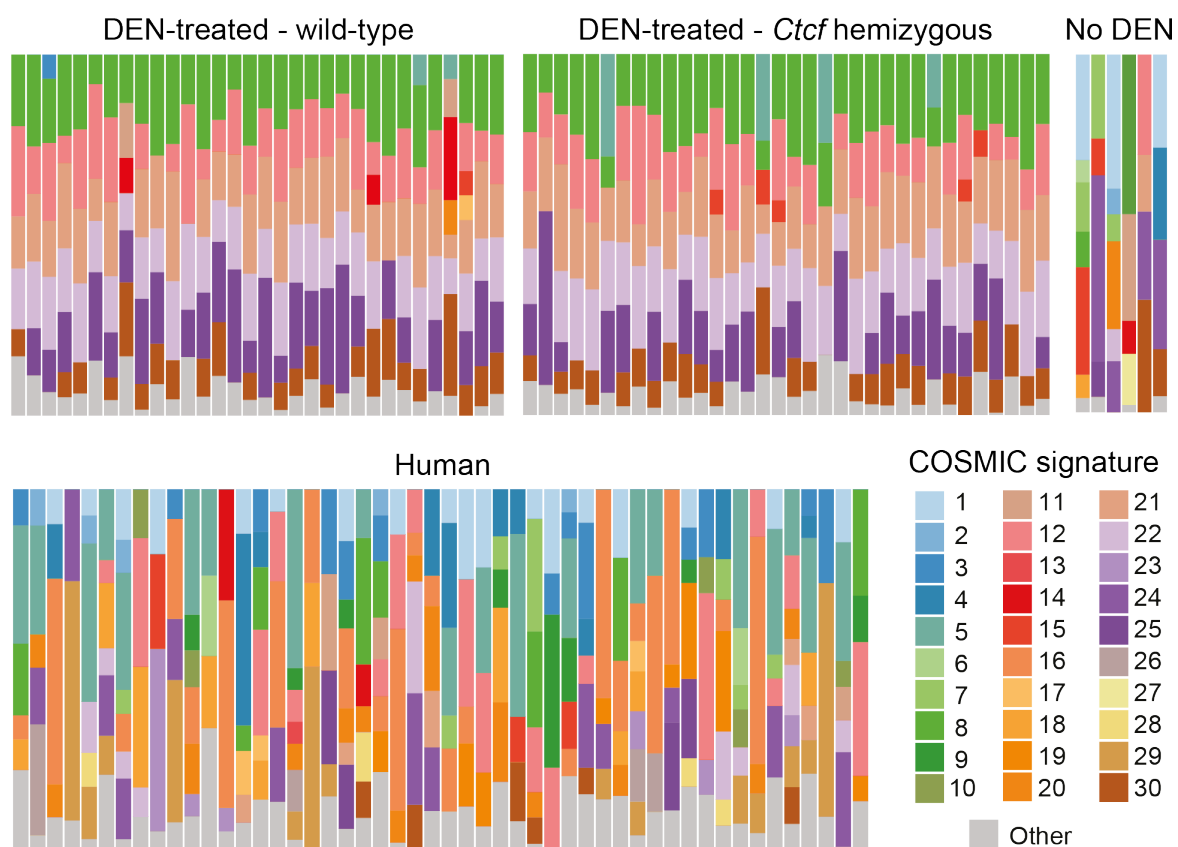


Fig. 4.11 Mutational signatures of DEN-induced and spontaneous tumours. Mutational portraits of individual mouse and human liver tumours reconstructed using COSMIC mutational signatures. Each column shows the composition of signatures in an individual sample. DEN-induced tumours from both wild-type and *Ctcf* hemizygous mice showed reproducible portraits largely composed of six component COSMIC signatures. In contrast, spontaneous tumours in mice and HCCs arising in the human cohort were more heterogeneous.

4.2.7 *Braf* is the predominant driver of DEN-induced hepatocarcinogenesis

All carcinogen-induced tumours carried a high burden of somatic mutations due to the known genotoxic action of DEN. These mutations were widely distributed across the genome of tumours arising in both *Ctcf* hemizygous and wild-type mice. Potential coding changes were detected in the majority (15,093) of protein-coding genes across the complete dataset, of which 7,257 were common to both genotypes. Identifying *bona fide* cancer drivers in the context of such a dense background mutational load is challenging and requires the combination of tractable experimental

models (Bakiri and Wagner, 2013) and the application of advanced computational biology approaches (Martincorena et al., 2017).

In order to identify putative driver genes in this cohort of DEN-initiated tumours, we identified cancer genes (Vogelstein et al., 2013) carrying non-synonymous protein-coding mutations more frequently than expected (**Figure 4.12**), with additional weight being given to genes with recurrent hotspot SNVs (**Table 4.5**).

Braf was the most highly mutated gene in carcinogen-induced tumours and was mutated in 58.8% of *Ctcf* hemizygous (20/34) and 59.4% of wild-type tumours (19/32; two-tailed Fisher's exact test, $p = 1$). *Hras* was also significantly mutated in tumours from mice with both genotypes; in 17.6% of *Ctcf* hemizygous (6/34) and 15.6% of wild-type (5/32) tumours (two-tailed Fisher's exact test, $p = 1$) (**Appendix B**). The predominant activating *Braf* mutation in both *Ctcf* genotypes was caused by an A:T to T:A transversion at the second base of codon 637 in *Braf*, which causes a valine to glutamic acid substitution (V637E), in keeping with previous studies in C57BL/6 mice (Buchmann et al., 2008). The mouse *Braf* V637E mutation is orthologous to the human *BRAF* V600E mutation (Rad et al., 2013), which is frequently detected in certain human cancers including approximately 50% of melanomas (Ascierto et al., 2012). Although rarer, the *Hras* mutations occurred almost exclusively at one hotspot, codon 61, most commonly a glutamine to arginine substitution caused by an A:T to G:C transition in the second base, consistent with the formation of one of the major promutagenic adducts, O⁴-ethyl-thymine, by DEN metabolites. In the more hepatocarcinogenesis-susceptible C3H mice, we found that these *Hras* codon 61 mutations were more prevalent than *Braf* mutations (Connor et al., 2018).

Spontaneously arising tumours had a much lower mutational burden, most frequently occurring in *Ctnnb1* (**Figure 4.12, Table 4.5**). A putative driver mutation could not be identified for two tumours, reflecting the polyclonal composition of the spontaneous tumours and/or involvement of other genetic or epigenetic alterations during tumourigenesis.

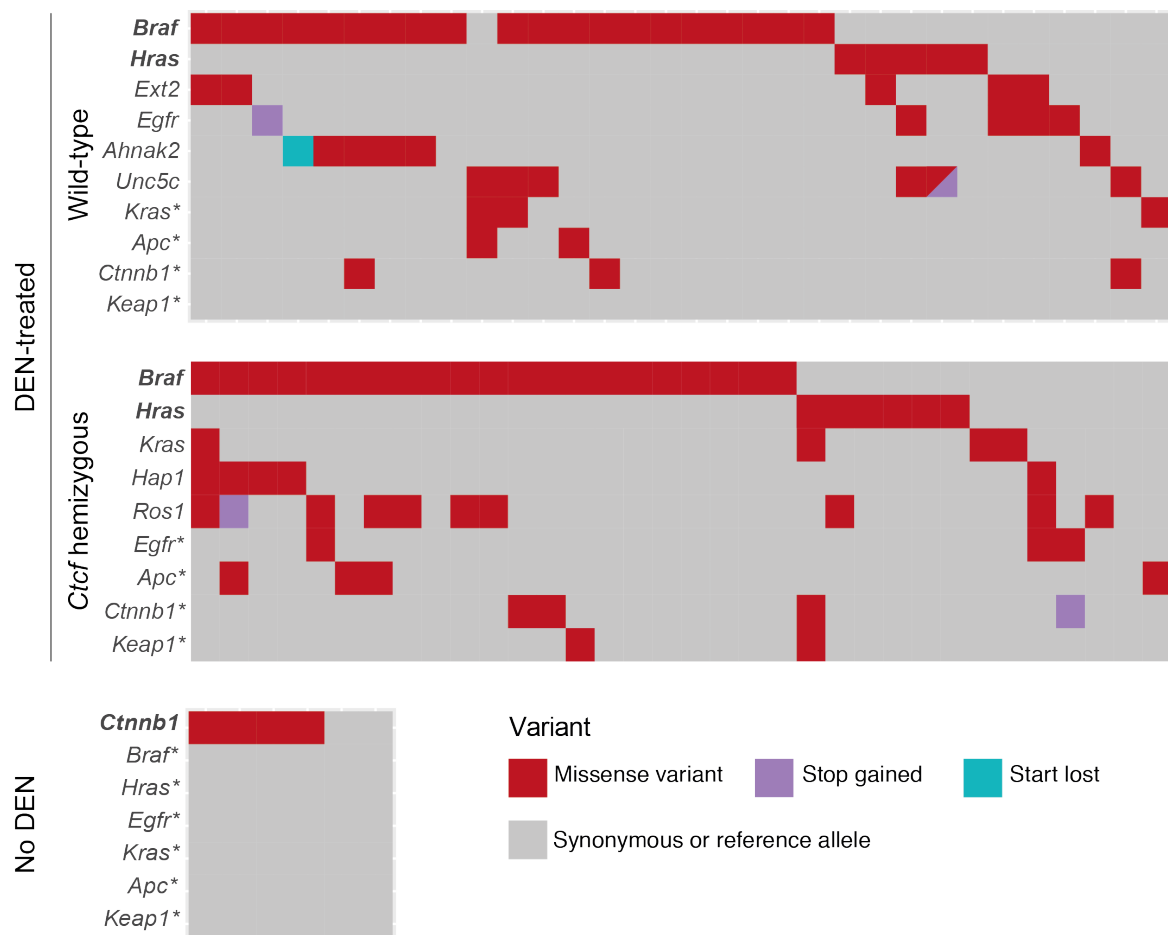


Fig. 4.12 DEN-initiated liver tumours carry *Braf* and *Hras* mutations. DEN-initiated liver tumours carry recurrent activating mutations in *Braf* and *Hras* in both *Ctcf* hemizygous and wild-type mice. In addition, there is diversity in non-synonymous SNVs in many other cancer genes. Each column represents a mouse tumour sample and each row a gene showing the occurrence of somatic variants in individual samples. Genes mutated in at least two samples are shown. All plots include the top seven genes mutated in other rodent tumours: *Hras*, *Braf*, *Egfr*, *Kras*, *Apc*, *Ctnnb1*, *Keap1* (unpublished data; indicated with * if not significant).

Table 4.5 Hotspot mutations in validated oncogenes and tumour suppressor genes. The frequency of tumours with non-synonymous SNVs in putative driver genes is given for each mouse cohort.

Driver gene	Mutation hotspot	DEN-treated mice		Spontaneous tumours
		Wild-type	<i>Ctcf</i> ^{+/-}	
<i>Braf</i>	V637E	19	19	0
	E205G	0	1	0
<i>Hras</i>	Q61R	4	3	0
	Q61K	1	2	0
	I24N	0	1	0
<i>Egfr</i>	F254I	3	2	0
	F150I	1	0	0
	I541N	0	1	0
	C506R	1	0	0
	M827T	1	0	0
	L885*	1	0	0
	M1009K	1	0	0
<i>Kras</i>	Q61R	2	4	0
	I21T	1	0	0
<i>Ctnnb1</i>	T41A	1	0	1
	S33F	0	0	1
	G34R	0	1	0
	S45F	0	0	2
	L46Q	0	0	1
	N121K	0	1	0
	I140T	0	1	0
	C213*	0	1	0
	W383R	1	0	0
	V438E	1	0	0
<i>Apc</i>	I402N	1	1	0
	T928I	1	0	0
	Y1029N	0	1	0
	S1200P	0	1	0
	C1725S	0	1	0
<i>Keap1</i>	M456V	0	1	0
	C273R	0	1	0

4.2.8 *Apc* is a secondary driver in mouse HCC

Given the high frequency of potentially consequential non-synonymous somatic variants, including in known oncogenes and tumour suppressor genes, the identification of secondary drivers was challenging. However, WGS of the thirteen DEN-initiated HCCs revealed that six of these had missense mutations in *Apc* and an additional six harboured β -catenin mutations (**Figure 4.12**). In contrast, these mutations were not found in pre-invasive dysplastic nodules. Furthermore, two thirds (4/6) of the spontaneously arising tumours had missense β -catenin mutations. This is comparable to human HCC cohorts in which the WNT/ β -catenin pathway is disrupted in 54% of patients (Schulze et al., 2015).

The highly conserved WNT/ β -catenin pathway is complex and can be perturbed in a number of ways (Chien et al., 2009). β -catenin protein, encoded by *CTNNB1*, is physiologically present in cells in three distinct pools: at cellular adherens junctions, in the cytoplasm, and in the nucleus (White et al., 2012). Cytoplasmic and nuclear β -catenin are maintained at low baseline levels by cytoplasmic phosphorylation of the N-terminus of β -catenin by a complex of proteins including APC (Yost et al., 1996). However, this process is imbalanced by canonical Wnt signalling, which prevents the degradation of β -catenin and causing it to accumulate in the cytoplasm, translocate to the nucleus, and in turn *trans*-activate target genes responsible for diverse processes including differentiation, proliferation, migration, and adhesion (White et al., 2012).

In order to assess whether the *Apc* mutations in our tumour cohort disrupted the Wnt/ β -catenin pathway, IHC was performed on tissue sections using antibodies targeting β -catenin. All *Apc*-mutant tumours showed aberrantly elevated nuclear β -catenin (**Figure 4.13**), while, as expected, the same IHC on DNs and HCCs with wild-type *Apc* and *Ctnnb1* showed normal membranous staining. These data support the hypothesis that mutant *Apc* disrupts the canonical Wnt/ β -catenin pathway to play a role in the progression to carcinoma in this model.

4.2.9 CTCF binding sites are enriched for mutations

Human cancer studies have highlighted that the somatic mutation rate is unevenly distributed across the genome (Lawrence et al., 2013) and is associated with transcriptional activity, the chromatin state, replication timing, and nuclear topology. In particular, regulatory sites show elevated substitution rates in cancers (Kaiser et al., 2016), most notably CTCF/cohesin-binding sites (Katainen et al., 2015; Sabari-

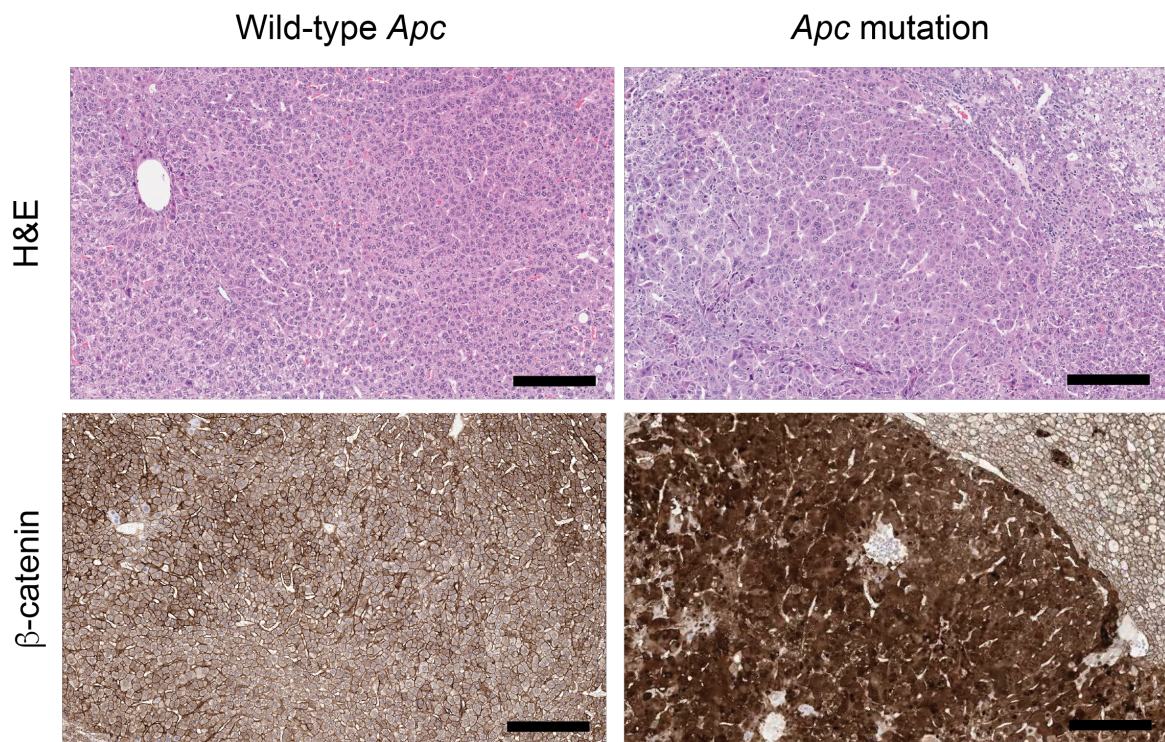


Fig. 4.13 Immunohistochemical validation of *Apc* mutations. Representative photomicrographs of serial tissue sections of DEN-induced HCCs. H&E staining demonstrates similar tumour morphology in tumours with wild-type *Apc* (left panels) and *Apc* mutations (right panels). IHC using antibodies targeting β -catenin protein demonstrates aberrantly elevated nuclear β -catenin protein expression in tumours with mutant *Apc*. All scale bars = 200 μ m. Original magnification x100. These data are presented in a similar form in Connor et al. (2018)

nathan et al., 2016). The underlying mechanisms are not fully understood (Flavahan et al., 2017; Stricker et al., 2016), although the high mutational burden observed in protein-bound DNA regions of melanomas is proposed to be due to impaired nucleotide excision repair (Sabarinathan et al., 2016).

Whole genome sequencing of our mouse liver tumour cohorts demonstrated a very high mutational burden that was non-uniformly distributed across the genome and that showed distinct mutational signatures, as described above. The distribution of mutations in tumours arising in both wild-type and *Ctcf* hemizygous mice recapitulated the features identified in previous analyses of human cancer cohorts (Frigola et al., 2017; Liu et al., 2013a; Park et al., 2012). The mutation rate was lowest in early replication timing regions of the genome and highest in late replication timing

regions, the SNV rate was anti-correlated with gene expression levels, and the SNV rate was lower in exons compared to introns.

Next, we focussed on mutations arising in the immediate vicinity of CTCF binding sites. All CTCF binding sites across all tumours of the same genotype were centred on their core CTCF motif, and mutations were quantified 400 bp upstream and downstream of these regions. This analysis showed that the mutation rate was elevated at the centre of the CTCF motif compared with its flanking regions (**Figure 4.15**). The frequency of observed mutations was significantly greater than the expected probability rate of mutations based on their trinucleotide context. Within the flanking regions, there was an oscillating distribution of mutations with a periodicity that correlated well with nucleosome positioning. There was no significant difference between the mutation rate or distribution between the two *Ctcf* genotypes. Next, using our CTCF ChIP-seq data from liver tissue, all CTCF binding sites were grouped into quartiles according to their affinity score and the number of somatic variants at each of these sites quantified. Higher affinity CTCF binding sites also accumulated a greater number of SNVs in both genotypes (**Figure 4.14**).

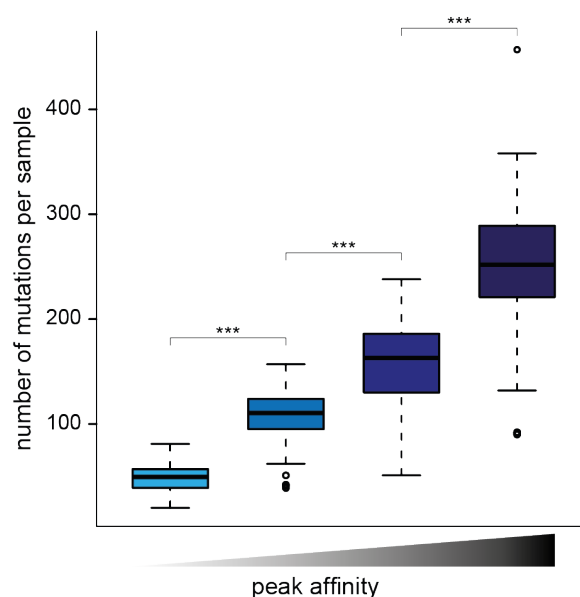


Fig. 4.14 CTCF binding affinity correlates with mutational burden. CTCF sites identified from ChIP-seq data in liver tissue were binned in quartiles according to number of read counts, and mutational burden was quantified for each site. Sites with stronger CTCF binding harbour more mutations than those with weaker CTCF binding. This is consistent with Sabarinathan et al. (2016), who showed that DNA motifs with high-affinity binding of transcription factors have a lower rate of DNA repair and thus accumulate more mutations.

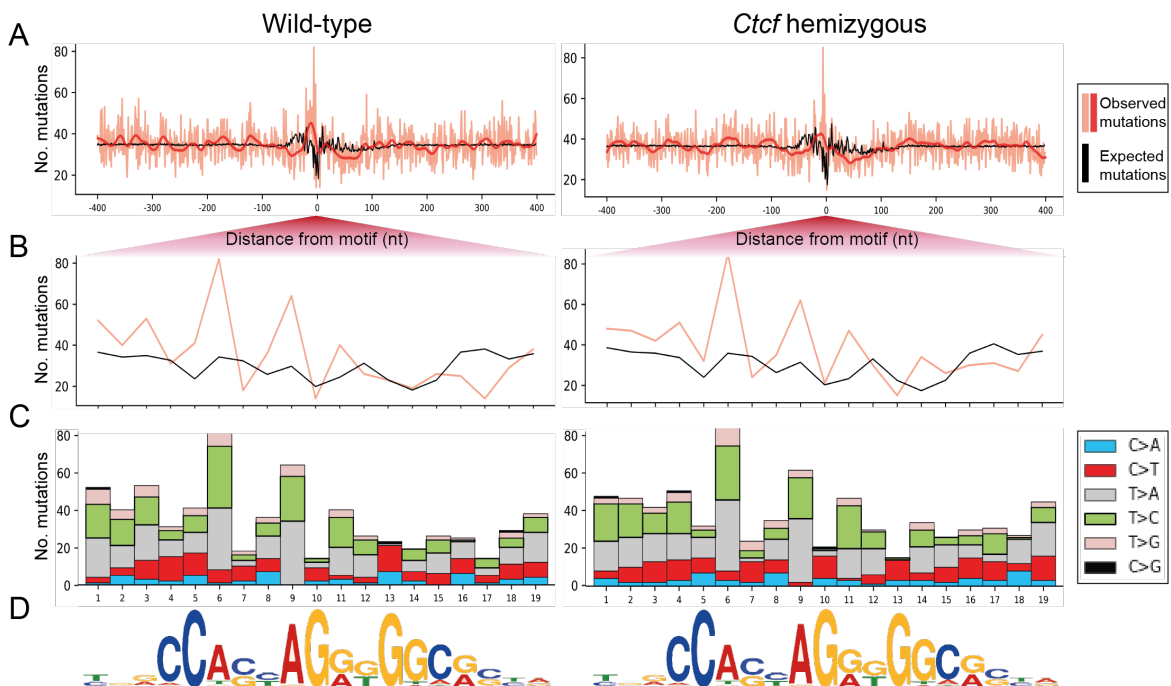


Fig. 4.15 CTCF binding sites are hypermutated. (A) The total (pale red) and average (dark red) number of observed mutations are shown at CTCF binding sites +/- 400 nt. The predicted mutation rate, based on trinucleotide context, is shown in black. Mutation rate is elevated at the CTCF motif compared with the flanking sites. There is a periodicity to the mutation rate in flanking regions which correlates with nucleosome positioning (DNA wrapped around a nucleosome is ~146 nt). (B-C) Mutation rate at each base within the core CTCF motif. (B) The observed frequency of mutations (red) varies across the CTCF motif and is not explained by the trinucleotide context-predicted mutation rate (black). (C) Nature of mutated nucleotides at each position of the CTCF motif. (D) The underlying CTCF motif.

4.2.10 *Carmil2* is overexpressed in *Ctf* hemizygous tumours

To characterise the transcriptional consequences of chemically inducing liver tumours in *Ctf* hemizygous mice, total RNA sequencing of the same tissues was undertaken. As expected, *Ctf* expression levels in the background liver tissue of DEN-exposed mice remained lower in hemizygous mice compared to their wild-type littermates (**Figure 4.16A**). Surprisingly, however, although all *Ctf* isoforms had lower expression in *Ctf*^{+/-} tumours compared to wild-type tumours, the difference was no longer significant (q value = 0.435). After correcting for multiple testing, only four genes were significantly differentially expressed between DEN-induced tumours of each genotype: *Carmil2* (log₂ fold change = 3.68), *Fbxo6* (log₂ fold change = 0.89), *Nudt11* (log₂ fold change = 2.79), *Mnd1* (log₂ fold change = 0.66,

Figure 4.16B). Of note, *Carmil2*, which is very significantly overexpressed in *Ctcf* hemizygous tumours, lies less than 3.5 kb downstream of *Ctcf* and is transcribed in the same direction. In these mice, coding exons 3 - 12 have been excised in one allele but exons 1 and 2 are still present (although noncoding). It is possible that the overexpression of *Carmil2* this is due to transcriptional read-through from the promoter of the genetically altered *Ctcf* locus.

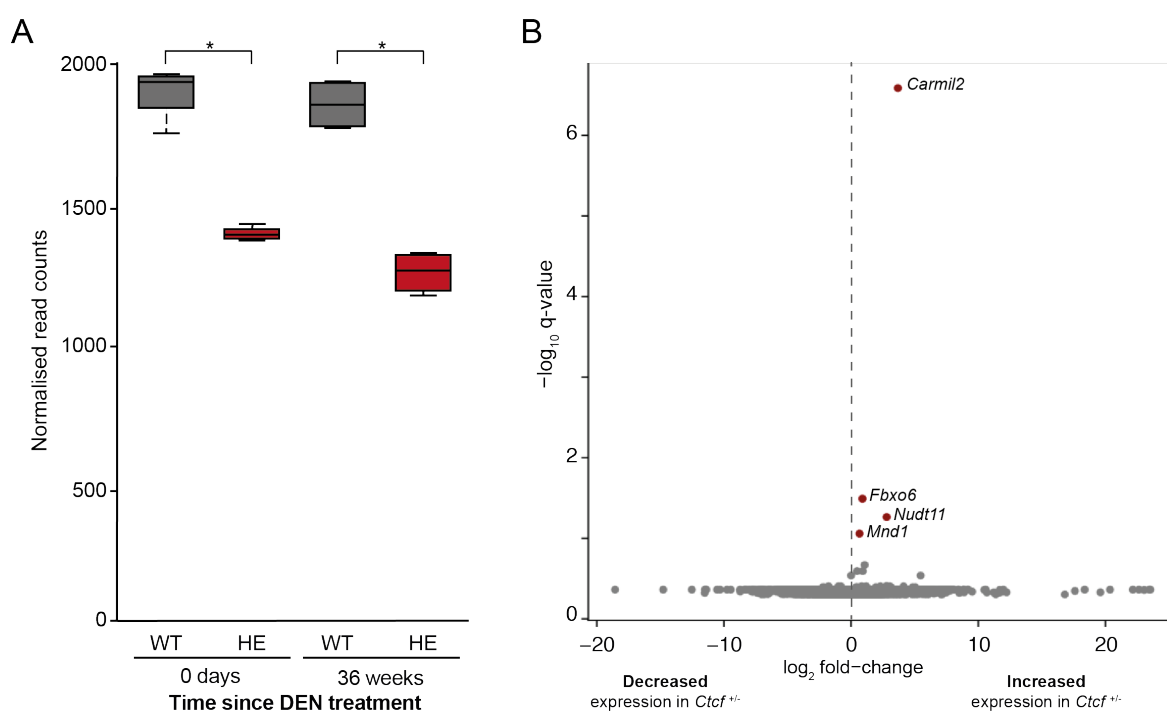


Fig. 4.16 Differential gene expression in liver and DEN-induced tumours. (A) *Ctcf* gene expression in liver tissue from wild-type (grey) and *Ctcf* hemizygous (red) 15-day-old mice (time 0) and non-tumour liver tissue 36 weeks after mice were treated with DEN. There is a significant reduction in *Ctcf* at both time points. (B) Differential gene expression analysis of total RNA-sequencing of DEN-induced tumours in *Ctcf* hemizygous versus wild-type mice revealed that only four genes are significantly differentially expressed (red): *Carmil2*, *Fbxo6*, *Nudt11*, and *Mnd1*, as indicated. \log_2 fold change of median expression differences are plotted against $-\log_{10}$ q value.

4.3 Discussion

CTCF is a principal nuclear protein that is essential for chromatin organisation and necessary for embryonic development. We and others (Heath et al., 2008) have shown that genetically-engineered mice with a germline deletion of one *Ctcf* allele have reduced transmission and/or survival of *Ctcf* hemizygous embryos compared to wild-type littermates. Surviving *Ctcf* hemizygous mice did not display any overt post-natal developmental defects: pups thrived, weaned successfully, and matured normally despite having lower CTCF expression. We also generated hepatocyte-specific *Ctcf* hemizygous mice and, in this instance, there was no detrimental impact on survival of the offspring. These mice were challenged with a potent carcinogen and, although there was no influence of the genotype on tumourigenesis, hemizygous mice showed a subtle phenotype of higher body and liver weights than their wild-type littermates after ageing.

In contrast to prior findings that CTCF haploinsufficiency predisposes to cancer (Kemp et al., 2014), our mouse cohort with hemizygous deletion of *Ctcf* did not have an increased incidence of spontaneous neoplasms compared to their wild-type littermates. However, far fewer tumours were found overall in our cohort compared with the Kemp et al. study. There are at least three possible explanations for these differences. First, Kemp et al. performed a survival study and, as such, were able to generate Kaplan Meier curves extending to later timepoints than in our study. Several of our mice developed health concerns at 18 - 20 months that meant that we had to use this as our fixed end point (due to Procedure Project Licence constraints). It is likely that, had we been able to extend our study, the tumour incidence in the population as a whole would have been higher and it is possible that there would have been differences between the genotypes. Second, our study was performed using C57BL/6J mice (the *Ctcf* deletion was originally engineered in a hybrid C57BL6/129 mouse but has now been back-crossed on a C57BL/6J background for over twenty generations), whereas the Kemp et al. study was carried out using C57BL6/129 F1 *Ctcf*^{+/-} mice. C57BL/6 mice are refractory to many tumour types (Puccini et al., 2013), so the relatively more tumour-resistant genetic background used here may have delayed a tumourigenesis phenotype beyond the duration of our study or masked a more subtle phenotype. Finally, we only included male mice in our study whereas Kemp et al. used both male and female mice. Overall, liver tumour incidence is three-times higher in male mice than female mice, and this is reflected in the findings

of these two studies. However, many of the tumours in Kemp's study were uterine tumours, which were of course not present in our male cohort.

The *Ctcf* hemizygosity model was challenged with DEN, a potent genotoxin. Such chemical carcinogen-initiated mouse models of liver cancer are important tools that are widely used to study the molecular pathogenesis of human HCC (Heindryckx et al., 2009). Genomic and transcriptomic studies have provided detailed characterisation of the aberrations found in the human disease (Ally et al., 2017; Fujimoto et al., 2016; Letouzé et al., 2017; Schulze et al., 2015, 2016; Zucman-Rossi et al., 2015), and it is important to have similar analyses from mouse models to inform human disease studies (Nassar et al., 2015; Westcott et al., 2014). Mice of both genotypes, *Ctcf*^{+/-} and *Ctcf*^{+/+}, developed multiple liver tumours by 36 weeks, and the frequency and size of tumours was independent of the mouse genotype. Of note, liver-specific *Ctcf* hemizygous mice aged for 42 weeks were, on average, 3.1 g heavier than their wild-type littermates, and the liver weights were also higher in the hemizygous mice. Tumour burden (both number and size of tumours) was equal in each genotype, suggesting that there was a genuine increase in the parenchymal liver weight. Increased liver weight accounted for over a third of the total body weight difference (1.2 g) between genotypes. Furthermore, at necropsy, *Ctcf* hemizygous mice had a pale "nutmeg" appearance to the visceral surface of the liver and a marked increase in intra-abdominal visceral fat. These subjective observations were not quantified in this series of experiments due to difficulties in applying a histochemical stain (oil red-O) to quantify fat. However, these differences - suggestive of fatty liver disease - could be due to hepatocyte-specific *Ctcf* hemizygosity causing either (i) steatosis resulting in fatty liver disease and subsequent weight gain due to dysregulated fat metabolism, or (ii) increased appetite/food consumption in turn leading to steatosis.

Whole genome sequencing revealed that, overall, the distribution of mutations in this carcinogen-initiated mouse model was consistent with results from human cancer sequencing studies. The SNV frequency was significantly higher in late DNA replication timing regions (Liu et al., 2013b), anti-correlated with gene expression levels (Park et al., 2012), and lower in exons compared with introns (Frigola et al., 2017). We confirmed that *Braf* is the predominant, although not obligatory, oncogenic driver of hepatocarcinogenesis in 15-day-old male C57BL/6J mice administered a single dose of DEN. *Hras* also harboured recurrent mutations, and many other oncogenes and tumour suppressor genes contained more sporadic SNVs. *Apc* was implicated as a secondary driver mutation since it was mutated in HCCs but

not DNAs, and these *Apc*-mutant tumours expressed aberrant levels of nuclear β -catenin. This is noteworthy because the WNT/ β -catenin pathway is altered in 54% of human HCCs, second only to *TERT* promoter mutations activating telomerase expression (Schulze et al., 2015). Spontaneous and DEN-induced tumours were macroscopically and microscopically indistinguishable but showed very different molecular signatures. Although the DEN model of tumourigenesis does not perfectly mimic the molecular portraits of the human disease, it may be a useful model in certain scenarios. In particular, this genomic study highlights the importance of undertaking molecular characterisation of mouse models in order to select the optimal model for a given scientific question. For example, pre-clinical models to test the efficacy of targeted therapeutics may rely on the presence (or absence) of specific mutations or expression profiles and will therefore need to be tested for prior to conducting the experiments.

Integrating the WGS and CTCF ChIP-seq datasets demonstrated that CTCF binding sites were enriched for mutations at their core motif and that higher affinity CTCF binding sites accumulated a disproportionate mutational burden. These findings are concordant with those seen in human melanoma samples. Sabarinathan et al. (2016) showed that the somatic mutation rate was highly increased at active transcription factor binding sites and nucleosome embedded DNA compared with their flanking regions. This important finding supports that this is a more generalisable feature of cancer genomes rather than being specific to a particular tumour type and/or its initiation and/or repair mechanism, e.g., UV light in the case of human melanoma or DEN in our mouse liver tumours.

RNA-sequencing of the tumours revealed that *Ctcf* expression levels in the background liver tissue of DEN-exposed mice remained lower in hemizygous mice than their wild-type littermates. Surprisingly, *Ctcf* expression in liver tumours was not significantly different between the two genotypes. One possible explanation for this could be the polyploid state of the liver. Unlike most mammalian tissues, the adult liver is composed of largely polyploid hepatocytes (up to 90% in adult mice) (Duncan et al., 2010). Polyploid hepatocytes arise due to cytokinetic failure (resulting in binucleated hepatocytes (Margall-Ducos et al., 2007)) or endoreduplication (replication of the nuclear genome in the absence of cell division) (Gentric and Desdouets, 2014). Tetraploid or octaploid hepatocytes can, therefore, be bi- or mononuclear. In the context of a hemizygous allele in the original diploid cell, selective pressure may mean that the mutant allele becomes enriched or depleted over time. This polyploid state is thought to serve an important tumour suppressive role in the liver, and it

has been shown experimentally that polyploid livers are protected against tumour suppressor loss of heterozygosity (Zhang et al., 2018). However, this is unlikely to have occurred in our mice since the majority of somatic variants were caused by an initial single burst of mutagenesis in the originating cell upon DEN exposure at 15 days, whereas liver ploidy in rodents only begins around weaning (postnatal day 14 to 21) and increases with ageing (Margall-Ducos et al., 2007; Schwartz-Arad et al., 1989). However, the presence or absence of this phenomenon could be investigated by performing DNA-FISH in tumour tissue sections using probes against wild-type *Ctcf* and the excised region of *Ctcf*, relative to the total count of chromosome 8.

Analysis of the tumour RNA-seq data revealed that only four genes were differentially expressed between tumours arising in *Ctcf* hemizygous and wild-type mice: *Carmil2*, *Fbxo6*, *Nudt11*, and *Mnd1*. Most significantly, *Carmil2* was 13.5-fold overexpressed (q value = 2.58×10^{-7}) in *Ctcf* hemizygous mice. *Carmil2* is a cell membrane-cytoskeleton-associated protein that plays roles in the regulation of actin polymerisation at the barbed end of actin filaments (Liang et al., 2009), cell migration, and invadopodia formation (Lanier et al., 2015); the latter is associated with epithelial-mesenchymal transition (EMT) and metastasis in cancer (Murphy and Courtneidge, 2011). *Carmil2* is encoded immediately downstream of *Ctcf*, so it is possible that the observed overexpression was due to transcriptional read-through from a *Ctcf* promoter upregulated in an attempt to compensate for depleted CTCF levels.

Of the other three genes, the most relevant to tumorigenesis is *Fbxo6*, which is involved in DNA damage responses by specifically recognising activated CHEK1 (phosphorylated Ser-345), promoting its ubiquitination and degradation. Ubiquitination of CHEK1 is required to ensure that activated CHEK1 does not accumulate as cells progress through S phase or when replication forks encounter transient impediments during normal DNA replication (Yoshida et al., 2005; Zhang et al., 2009). *Mnd1* is required for homologous chromosome pairing and efficient cross-over and intragenic recombination during meiosis (Petukhova et al., 2005). *Nudt11* is involved in signal transduction by cleaving beta-phosphate from diphosphate groups, and it also catalyses the hydrolysis of dinucleoside oligophosphates (Fisher et al., 2002). An important future direction is to use the RNA-seq data to validate the functional impact of putative drivers identified in WGS analyses.

In this study we used a genetically engineered mouse model to study spontaneous tumorigenesis with a reduced germline genomic concentration of *Ctcf*. Further, we generated a tissue-specific model of CTCF depletion and used chemical

carcinogenesis to challenge this further to characterise the resulting phenotypes. In contrast to previously published findings, mice hemizygous for *Ctcf* either at the whole organism level or in a tissue specific manner did not have an increased propensity to develop spontaneous or carcinogen-initiated tumours. However, these experiments provide a detailed histological, genomic, and transcriptomic characterisation of a cohort of liver tumours in C57BL/6J mice induced using a commonly used mouse model.

Chapter 5

Discussion and outlook

5.1 CTCF haploinsufficiency

CTCF is a pleiotropic DNA-binding protein with *in vitro* and *in vivo* evidence of a tumour suppressor role. Complete *Ctcf* knockout in mice is non-viable, and *Ctcf* hemizygous mice have been reported to be susceptible to spontaneous, radiation-, and chemically-induced tumours of multiple lineages. This series of experiments tested the hypothesis that a reduced genomic concentration of *Ctcf* increases susceptibility to cancer by altering chromatin homeostasis. We aimed to directly test this by performing a systems-level characterisation of the cells and tissues of *Ctcf* hemizygous mice and subsequently inducing and profiling liver tumours to identify (epi)genomic features predicting hypersensitive loci. Reducing the concentration of CTCF was expected to have a quantitative effect on DNA binding, either a reduced number of occupied binding sites or a partial global reduction in binding at all sites. Furthermore, we expected the downstream effects on gene transcription and protein expression would be dependent on their relationship to altered sites of activation or repression and/or altered CTCF-dependent chromatin looping. We also predicted that liver tumourigenesis would be accelerated in *Ctcf* conditional knockout mice.

We found that *Ctcf* hemizygous mammalian cells exquisitely compensate for fluctuations in intra-cellular CTCF concentration. *In vitro* removal of 50% of the *Ctcf* gene content caused a 37% reduction in mRNA expression and a 27% reduction in CTCF protein levels, which resulted in only a 2% difference in genomic occupancy by CTCF. In spite of this degree of compensation, genetic and epigenetic profiling of *Ctcf* hemizygous MEF cultures revealed reproducible changes in the nuclear environment, thus providing insights into the inherent functions of this essential

protein. Several hundred genes and their corresponding proteins were differentially expressed and were enriched for cancer-related pathways. The effect sizes were not as great as we hypothesised, although are in keeping with Zuin et al. (2014), who found only 161 differentially expressed genes in *CTCF* RNAi experiments.

Induction of liver tumours with a carcinogen provided detailed insights into the mutational signatures and transcriptional profiles of DEN-initiated tumours in C57BL/6J mice. Tumourigenesis was not accelerated in *Ctcf* conditional knockout mice, and tumours arising in wild-type and *Ctcf* hemizygous mice were only subtly phenotypically different. However, this WGS dataset represents an important resource to study the effects of controlled mutagenesis *in vivo*. Future directions in this regard are described below.

Although we adopted a high-throughput approach to examine CTCF loss *in vitro* and *in vivo*, future studies could adopt a locus-centric approach to complement these data and functionally test the consequences of CTCF depletion or the loss of CTCF binding. For example, CRISPR techniques could be used to delete specific CTCF binding sites. By studying the presence (or absence) of transcriptional and chromatin conformational changes, the functional role for transcriptional or structural homeostasis could be characterised. This has proven to be a powerful way to dissect the roles of individual regulatory elements of the α -globin super-enhancer (Hay et al., 2016) and proving the insulator role of CTCF in *IDH1*-mutant gliomas (Flavahan et al., 2015).

5.2 Model systems

Model systems provide powerful experimental tools to study fundamental biological processes and address particular biological questions. There is of course a wide selection of model organisms to choose from including bacteria (e.g., *Escherichia coli*) and bacteriophage viruses, bakers' yeast (*Saccharomyces cerevisiae*), nematode worms (*Caenorhabditis elegans*), fruit flies (*Drosophila melanogaster*), zebrafish (*Danio rerio*), frogs (*Xenopus laevis*), and mice (*Mus musculus*). The choice of model depends upon the question being asked and the suitability of the organism for answering that question (Fields and Johnston, 2005). For instance, a mammalian model is far more applicable for pharmacokinetic and pharmacodynamic studies on new drugs than, say, the fruit fly or worm, since to test these parameters a whole organism with a blood circulation, liver, and kidneys is needed. In this study we used mice and derived cell lines to study CTCF in a mammalian system and to induce

tumours. We also took advantage of Cre recombinase technology to genetically engineer the *Ctcf* locus in order to manipulate its expression in the germline and in a conditional, cell-type specific manner. While this represented a sophisticated set of approaches to test our hypotheses, these models (like any experimental model) had their limitations.

Experiments in MEFs were performed in bulk culture of millions of cells. There may well be more cell-to-cell variability in the *Ctcf* hemizygous state, but the nature of this heterogeneous and unsynchronised cell culture may mean that quantitatively large changes in some cells were masked. Single-cell sequencing technologies (Schwartzman and Tanay, 2015; Wang and Navin, 2015; Wu et al., 2017) are rapidly becoming more accessible and less expensive, and represent one way to examine individual cell variability. For example, single-cell RNA-seq (scRNA-seq) can be performed by fluorescence activated cell sorting (FACS) and then capture of hundreds of individual cells for library preparation, for example with Smart-seq2, followed by sequencing. Alternatively, more recent droplet-based RNA-seq (drop-seq) technologies allow thousands of individual cells to be profiled in parallel, with drop-seq having the advantage that biases introduced by FACS sorting and capturing are negated (Habib et al., 2017; Klein et al., 2015; Macosko et al., 2015). However, this latter technology has relatively low gene-per-cell sensitivity compared to other scRNA-Seq methods (Ziegenhain et al., 2017). It is also possible to parallel sequence the genome and transcriptome of single cells (G&T-seq) (Macaulay et al., 2016) or to simultaneously profile chromatin accessibility, DNA methylation, and transcription in single cells (scNMT-seq) (Clark et al., 2018). In addition, the dynamics of chromosomal organisation can now be performed at single-cell resolution (Nagano et al., 2017). Although these technologies were not available during the course of this project, they represent exciting new ways to interrogate *in vitro* cultures and will undoubtedly reveal more complexity within the cell population as a whole than previously appreciated.

The *Ctcf* hemizygous mice were not born with equal numbers of wild-type and *Ctcf* hemizygous offspring as expected by Mendelian inheritance. This discrepancy in fertilisation, embryogenesis, and/or full-term development may have introduced bias into downstream experiments. While the reduced transmission rate (30.6%) is consistent with the genotypic ratios observed in previous studies using a C57BL/6 background (Heath et al., 2008), it is in contrast to the findings of Kemp et al. (2014) who reported that "C57BL6/129 (B6/129) F1 *Ctcf*^{+/-} mice were born at the expected Mendelian frequency". The reduced transmission rate may be explained by the

hybrid 129sv background used by Kemp et al. (2014), since breeding transgenic animals in F1 animals has been shown to be more efficient than a homogeneous C57BL/6 background (Owenab et al., 1997; Taketo et al., 1991). Nevertheless, important CTCF-mediated effects may have been missed if the reason for offspring imbalance was due to the effects of *Ctcf* hemizygosity *in utero*.

Strain-specific differences may also explain the reduced rate of spontaneous tumours in our CB57BL/6J animals compared with BL6/129 since the 129 family is particularly diverse, with numerous substrains across four separate genetic lineages containing known phenotypic differences (Kiselycznyk and Holmes, 2011). Although the reported overall tumour incidence in 129 mice was relatively low (7% in males, 21% in females), it was higher than in CB57BL/6J animals, which are refractory to many tumour types (Smith et al., 1973). The most frequently reported tumour types in 129 mice are lymphoma, soft tissue sarcomas, lung tumours, and testicular teratomas (Smith et al., 1973), which reflects those seen by Kemp et al. (2014). We also know from our own mouse colonies that tumour latency - both spontaneous and induced - varies widely between inbred laboratory strains (including C57BL/6J and C3H (Connor et al., 2018)) and wild-derived strains (including *Mus musculus castaneus* and *Mus caroli* (unpublished data)).

We also used a combination of hepatocyte-specific *Ctcf* deletion with a carcinogen tumour-induction protocol. We generated a detailed genomic and transcriptomic characterisation of this model but did not identify marked differences between *Ctcf* hemizygous and wild-type mice. The similarity between genotypes is likely due to a combination of factors: the choice of liver as the target organ, and the carcinogen model used. A hepatocyte-specific model was selected for several valid reasons outlined in Chapter 4. To recap, the liver was chosen because: (i) the liver cell population was regarded as relatively homogeneous for these studies (i.e., hepatocyte-rich in the loosest terms); (ii) it is well characterised by previous studies in our laboratory (Odom et al., 2007; Schmidt et al., 2010a,b; Schwalie et al., 2013; Villar et al., 2014); (iii) we have significant experience with a liver tumour-induction protocol (Connor et al., 2018); (iv) the liver is easily identified at necropsy; and (v) a single mouse liver provides sufficient tissue for several assays. However, an alternative target organ might have better bridged the phenotype-genotype gap. *CTCF* has been identified as a mutational cancer driver gene in uterine corpus endometrioid carcinoma (Rubio-Perez et al., 2015), and uterine neoplasms were found in 40% of all female mice studied by Kemp et al. (2014). It would, therefore, be interesting to test the same hypothesis in uterine-specific *Ctcf* hemizygous mice. This could be achieved

using a mouse expressing the modified progesterone allele: *Pgr*^{tm1.1(cre)Shah} (*PR*^{Cre}), which harbours a Cre recombinase gene downstream of the progesterone receptor (*Pgr*) transcriptional stop codon. Spatiotemporal expression of *cre* therefore follows endogenous progesterone expression. This nuclear hormone receptor is expressed in the uterus but is also highly expressed in all other female reproductive organs (including ovaries, fallopian tubes, and mammary glands), as well as the pituitary gland and hypothalamus. Therefore, although heterozygous mice are viable, fertile, normal in size and do not display any gross physical or behavioural abnormalities, the results might be obscured or confounded by *Ctcf* deletion in other tissues.

In relation to the carcinogen model, we found that *Braf*(V637E) was the most common oncogenic mutation in both genotypes, consistent with previous findings in C57BL/6 mice (Bakiri and Wagner, 2013). This is in contrast to the *Hras* and *Egfr* mutations predominating in C3H mice (Connor et al., 2018). We now know that the mutational signatures caused by DEN are not only the same between the two *Ctcf* genotypes but that the signature is also the same in C3H (Connor et al., 2018), *Mus musculus castaneus*, and *Mus Caroli* mice (unpublished data). Thus, mouse strain is the strongest determinant of mutational drivers, whereas the mutational signature caused by DEN is universal and was unlikely to be significantly altered by *Ctcf* hemizygosity alone.

Other model systems continue to evolve and that might provide interesting options for future investigation of *Ctcf* and liver cancer. Organoid models now allow human (or mouse) physiology and disease to be recapitulated three dimensionally *in vitro*, even in tissues with complex structure and architecture such as the brain (Lancaster et al., 2013) or liver (Broutier et al., 2017; Huch et al., 2013). Organoid (or tumouroid) culture systems can capture and maintain many features of the original human tissue (or tumour) sample including morphological subtype, immunohistochemical profile, gene expression, and mutational profile. In addition to providing a model for fundamental biology and pathophysiology, such systems also present options for drug screening and validation of actionable therapeutic targets for personalised and precision medicine programmes (Huch and Koo, 2015).

5.3 Mechanisms of chromatin organisation

The importance of CTCF in the three-dimensional organisation of chromatin is well established and, although the precise mechanisms are not fully resolved, several important discoveries have recently been made.

There is an increasing body of *in vitro*, *in vivo*, and *in silico* evidence to support the loop extrusion model of chromatin folding (see Chapter 1, **Figure 1.4**). One of these advances relates to the interplay between loop extrusion of TADs and compartmental segregation of interphase chromatin. Very recent reports over the last few months support that although TADs are subunits within compartments, their structures are distinct, and that loop extrusion of TADs actually counteracts the formation of compartments (Nuebler et al., 2018). These two independent modes of chromosome organisation have been demonstrated by: (1) targeted degradation of CTCF from mouse embryonic stem cells, which caused disruption of TADs because loop extrusion was no longer insulated within specific domains, but local chromatin compaction was maintained to counteract compartmental segregation (Nora et al., 2017); (2) removing *Nipbl*, a cohesin loader, revealed the intrinsically segregated compartmental structure in the absence of its suppression by loop extrusion (Schwarzer et al., 2017); and (3) removing *Wapl*, a cohesin unloading factor, resulted in increased loop length and abundance due to prolonged occupancy of cohesin on the DNA to strengthen TADs and weaken compartmentalisation (Haarhuis et al., 2017). Nuebler et al. (2018) have since used polymer models of chromosomes to quantitatively reproduce these experimental findings.

These studies strongly support the process of loop extrusion for interphase chromosome folding and the essential complementary roles of CTCF and cohesin. However, there are some aspects that remain unresolved, since these studies do not directly address the mechanism of loop extrusion. The process of extrusion requires the protein complex to travel continuously for thousands of kilobases along the chromatin in spite of obstructing nucleosomes and DNA-binding proteins (Fudenberg et al., 2018). Another aspect briefly discussed in Chapter 4 is the energy expenditure required for the active process of extrusion. However, more recent estimates show that the energy burden of ATP consumption by loop-extruding cohesins in interphase is negligible compared with total ATP production in a mammalian cell (Terakawa et al., 2017) and therefore unlikely to be a limiting factor.

Loop extrusion is also proposed to be the mechanism of chromosomal folding and compaction in mitosis, with condensin (rather than cohesin) acting as the mechanochemical motor (Terakawa et al., 2017). Polymer simulations based on a combination of high-resolution imaging and Hi-C of synchronous cell cultures (including condensin depletion studies) showed that the interphase organisation of chromatin is rapidly lost in prophase and 60 kb loops are formed. These inner loops become nested within ~400 kb outer loops in prometaphase, which progressively

increase to ~12 Mb as the loop array extends in a helical "spiral-staircase" condensin scaffold (Gibcus et al., 2018).

This detailed description was followed almost immediately by the first real-time imaging of DNA loop extrusion by condensin. Time-lapse imaging unambiguously demonstrated the formation and progressive extension of DNA loops by condensin, thus providing clear and direct evidence for active loop extrusion. Direct visualisation showed that a single condensin complex extrudes tens of kilobase pairs of DNA at speeds of up to 1.5 kb per second (Ganji et al., 2018). An unexpected finding that had not been proposed by *in silico* experiments was that condensin-dependent loop extrusion is strictly asymmetrical: condensin anchors onto DNA and reels it in from only one side (Ganji et al., 2018).

Although cohesin-dependent looping has not been directly visualised in inter-phase chromatin, this powerful evidence for condensin-dependent loop extrusion in mitosis further supports the role of an active loop extrusion process for other aspects of genome organisation.

5.4 DNA damage and repair

DEN is a potent genotoxin that causes DNA damage by two parallel processes: first, as a DNA alkylating agent, DEN leads to the formation of mutagenic DNA adducts (Bakiri and Wagner, 2013). Second, its bioactivation by cytochrome P450 in centrilobular hepatocytes generates reactive oxygen species (ROS) that damage DNA, proteins, and lipids and kills hepatocytes (Qi et al., 2008). In turn, hepatotoxicity triggers an inflammatory response resulting in elevated expression of mitogens such as interleukin-6 (IL-6), which promote compensatory proliferation of surviving hepatocytes (Naugler et al., 2007). The alkylation adducts can be removed by the DNA repair gene O⁶-methylguanine-DNA methyltransferase (MGMT), which encodes O⁶-alkylguanine-DNA alkyltransferase (Pegg, 1990).

Genome-wide, we showed that DEN-initiated liver tumours possess a reproducible mutational signature that can be attributed to the initial treatment with DEN, whereas driver mutations are strain/species-dependent (Connor et al., 2018). This mutational signature is the consequence of several interacting processes: the intrinsic infidelity of the DNA replication machinery, exogenous (e.g., DEN) or endogenous (e.g., ROS) mutagenic exposures, enzymatic modification of DNA, and defective DNA repair (Alexandrov et al., 2013). The incidence of mutations was not evenly distributed across the genome and was in keeping with previous reports of a re-

duced mutation rate in exons. In melanoma, this is thought to be due to differential mismatch repair (Frigola et al., 2017), and oxidative damage is also reduced at promoters, exons, and termination sites but not introns (Poetsch et al., 2018).

We found that CTCF binding sites were hypermutated in tumours of both *Ctcf* genotypes. This has also been shown in human melanomas, which are almost entirely driven by UV mutagenesis, and it has been proposed that this is due to impaired nucleotide excision repair (NER) (Poulos et al., 2016) due to DNA-bound transcription factors interfering with the NER machinery (Sabarinathan et al., 2016). More generally, analysis of over a thousand cancer genomes across 14 cancer types found increased mutational density at gene promoters, which was linked to transcription initiation activity and impairment of NER (Perera et al., 2016).

Therefore, one future direction will be to examine the impact of DNA repair on hypermutated CTCF binding sites. DEN causes predictable mutations due to its intrinsic chemistry and our tumours harboured a specific DEN signature. However, the presence of hypermutated CTCF binding sites showed that the DEN mutagenic effect was not evenly distributed across the genome. It is unclear how this uneven distribution relates to the balance between damage and repair occurring either at the time of initial insult or over time. Experimental mapping of early DNA damage and/or repair immediately after treatment with DEN would help to dynamically map the evolution of mutations and thus better interpret the final mutational signature. We know that the promutagenic O⁶-ethyl deoxyguanosine adduct accumulates in zone 3 hepatocytes within four hours of treatment with DEN and DNA double-strand breaks (as seen by the rapid accumulation and elimination of phosphorylated histone H2AX) peaks over the same time frame (Connor et al., 2018). Therefore, in order to directly assess DNA damage and before repair has occurred, liver samples would have to be taken within the first few minutes to hours after pups were treated with DEN. Several new protocols have been developed to map DNA damage and/or repair *in vitro* and *ex vivo*, as follows.

In replicating cells, ribonucleotides are the most common non-canonical nucleotides incorporated into the genome, and they are removed by ribonucleotide excision repair initiated by RNase H2 cleavage. In the absence of RNase H2, such embedded ribonucleotides can be used to track DNA polymerase activity *in vivo* to study DNA replication and repair at single nucleotide resolution. Strand-specific genome-wide mapping of embedded ribonucleotides (emRiboSeq) can be performed using recombinant RNase H2 to selectively create ligatable 3'-hydroxyl groups, in contrast to alternative methods that use alkaline hydrolysis. Non-canonical bases

can be mapped by substituting RNase H2 with specific nicking endonucleases: endonuclease sequencing (EndoSeq) (Ding et al., 2015). Alternatively, AP-seq can be used to map apurinic sites and 8-oxo-7,8-dihydroguanine bases at ~300 bp resolution on a genome-wide scale (Poetsch et al., 2018). Such techniques allow genome-wide maps of alkylation damage, repair, and mutagenesis at single nucleotide resolution to reveal mechanisms of mutational heterogeneity (Mao et al., 2017) and may shed new light on why environmental exposures have different effects on cancer predisposition.

Publications

The following publications have arisen from work carried out during my PhD:

SJ Aitken*, X Ibarra-Soria*, E Kentepozidou, P Flicek, C Feig, JC Marioni, DT Odom. CTCF maintains regulatory homeostasis of cancer pathways.
Genome Biology. 2018;19(1):106. doi:10.1186/s13059-018-1484-3.

F Connor*, TF Rayner*, **SJ Aitken**, C Feig, M Lukk, J Santoyo-Lopez, DT Odom. Mutational landscape of a chemically-induced mouse model of liver cancer.
Journal of Hepatology. 2018;69(4):840-850. doi:10.1016/j.jhep.2018.06.009.

I Quiros-Gonzalez, MR Tomaszewski, **SJ Aitken**, L Ansel-Bollepalli, LA McDuffus, M Gill, L Hacker, J Brunner, SE Bohndiek. Optoacoustics delineates murine breast cancer models displaying angiogenesis and vascular mimicry.
British Journal of Cancer. 2018;118(8):1098-1106. doi:10.1038/s41416-018-0033-x.

C Ernst, J Pike, **SJ Aitken**, HK Long, N Eling, L Stojic, MC Ward, F Connor, TF Rayner, M Lukk, RJ Klose, C Kutter, DT Odom. Successful transmission and transcriptional deployment of a human chromosome via mouse male meiosis
eLife. 2016;5:e20235. doi:10.7554/eLife.20235.

* denotes co-first authorship

References

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferreira, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirkas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–95.
- Aitken, S. J., Ibarra-Soria, X., Kentepozidou, E., Flicek, P., Feig, C., Marioni, J. C., and Odom, D. T. (2018). CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biology*, 19(1):106.

- Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D. N., Nag, R., Newman, V., Nuhn, M., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S. P., Zadissa, A., Kostadima, M., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Cunningham, F., Yates, A., Zerbino, D. R., and Flicek, P. (2017). Ensembl 2017. *Nucleic Acids Research*, 45(D1):D635–D642.
- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014). *Molecular Biology of the Cell*. Garland Science, 6th edition.
- Aleksic, K., Lackner, C., Geigl, J. B., Schwarz, M., Auer, M., Ulz, P., Fischer, M., Trajanoski, Z., Otte, M., and Speicher, M. R. (2011). Evolution of genomic instability in diethylnitrosamine-induced hepatocarcinogenesis in mice. *Hepatology*, 53(3):895–904.
- Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–7.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjord, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jager, N., Jones, D. T. W., Knappskog, S., Kool, M., Lakhani, S. R., Lopez-Otin, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21.
- Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Heisler, L. E., Beck, T. A., Simpson, J. T., Tonon, L., Sertier, A.-S., Patch, A.-M., Jäger, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., Previti, C., Schmidt, S., Brors, B., Feuerbach, L., Heinold, M., Gröbner, S., Korshunov, A., Tarpey, P. S., Butler, A. P., Hinton, J., Jones, D., Menzies, A., Raine, K., Shepherd, R., Stebbings, L., Teague, J. W., Ribeca, P., Giner, F. C., Beltran, S., Raineri, E., Dabad, M., Heath, S. C., Gut, M., Denroche, R. E., Harding, N. J., Yamaguchi, T. N., Fujimoto, A., Nakagawa, H., Quesada, V., Valdés-Mas, R., Nakken, S., Vodák, D., Bower, L., Lynch, A. G., Anderson, C. L., Waddell, N., Pearson, J. V., Grimmond, S. M., Peto, M., Spellman, P., He, M., Kandoth, C., Lee, S., Zhang, J., Létourneau, L., Ma, S., Seth, S., Torrents, D., Xi, L., Wheeler, D. A., López-Otín, C., Campo, E., Campbell, P. J., Boutros, P. C., Puente, X. S., Gerhard, D. S., Pfister, S. M., McPherson, J. D., Hudson, T. J., Schlesner, M., Lichter, P.,

- Eils, R., Jones, D. T. W., and Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, 6:10001.
- Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.
- Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Cheung, D., Wong, T., Brooks, D., Robertson, A. G., Bowlby, R., Mungall, K., Sadeghi, S., Xi, L., Covington, K., Shinbrot, E., Wheeler, D. A., Gibbs, R. A., Donehower, L. A., Wang, L., Bowen, J., Gastier-Foster, J. M., Gerken, M., Helsel, C., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Gabriel, S. B., Meyerson, M., Cibulskis, C., Murray, B. A., Shih, J., Beroukhi, R., Cherniack, A. D., Schumacher, S. E., Saksena, G., Pedamallu, C. S., Chin, L., Getz, G., Noble, M., Zhang, H. H., Heiman, D., Cho, J., Gehlenborg, N., Saksena, G., Voet, D., Lin, P., Frazer, S., Defreitas, T., Meier, S., Lawrence, M., Kim, J., Creighton, C. J., Muzny, D., Doddapaneni, H., Hu, J., Wang, M., Morton, D., Korchina, V., Han, Y., Dinh, H., Lewis, L., Bellair, M., Liu, X., Santibanez, J., Glenn, R., Lee, S., Hale, W., Parker, J. S., Wilkerson, M. D., Hayes, D. N., Reynolds, S. M., Shmulevich, I., Zhang, W., Liu, Y., Iype, L., Makhoul, H., Torbenson, M. S., Kakar, S., Yeh, M. M., Jain, D., Kleiner, D. E., Jain, D., Dhanasekaran, R., El-Serag, H. B., Yim, S. Y., Weinstein, J. N., Mishra, L., Zhang, J. J., Akbani, R., Ling, S., Ju, Z., Su, X., Hegde, A. M., Mills, G. B., Lu, Y., Chen, J., Lee, J.-S., Sohn, B. H., Shim, J. J., Tong, P., Aburatani, H., Yamamoto, S., Tatsuno, K., Li, W., Xia, Z., Stransky, N., Seiser, E., Innocenti, F., Gao, J., Kundra, R., Zhang, H. H., Heins, Z., Ochoa, A., Sander, C., Ladanyi, M., Shen, R., Arora, A., Sanchez-Vega, F., Schultz, N., Kasaian, K., Radenbaugh, A., Bissig, K.-D., Moore, D. D., Totoki, Y., Nakamura, H., Shibata, T., Yau, C., Graim, K., Stuart, J., Haussler, D., Slagle, B. L., Ojesina, A. I., Katsonis, P., Koire, A., Lichtarge, O., Hsu, T.-K., Ferguson, M. L., Demchok, J. A., Felau, I., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J. J., Hutter, C. M., Sofia, H. J., Verhaak, R. G., Zheng, S., Lang, F., Chudamani, S., Liu, J., Lolla, L., Wu, Y., Naresh, R., Pihl, T., Sun, C., Wan, Y., Benz, C., Perou, A. H., Thorne, L. B., Boice, L., Huang, M., Rathmell, W. K., Noushmehr, H., Saggiaro, F. P., Tirapelli, D. P. d. C., Junior, C. G. C., Mente, E. D., Silva, O. d. C., Trevisan, F. A., Kang, K. J., Ahn, K. S., Giama, N. H., Moser, C. D., Giordano, T. J., Vinco, M., Welling, T. H., Crain, D., Curley, E., Gardner, J., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Kelley, R., Park, J.-W., Chandan, V. S., Roberts, L. R., Bathe, O. F., Hagedorn, C. H., Auman, J. T., O'Brien, D. R., Kocher, J.-P. A., Jones, C. D., Mieczkowski, P. A., Perou, C. M., Skelly, T., Tan, D., Veluvolu, U., Balu, S., Bodenheimer, T., Hoyle, A. P., Jefferys, S. R., Meng, S., Mose, L. E., Shi, Y., Simons, J. V., Soloway, M. G., Roach, J., Hoadley, K. A., Baylin, S. B., Shen, H., Hinoue, T., Bootwalla, M. S., Van Den Berg, D. J., Weisenberger, D. J., Lai, P. H., Holbrook, A., Berrios, M., and Laird, P. W. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, 169(7):1327–1341.

- Anway, M. D., Cupp, A. S., Uzumcu, M., and Skinner, M. K. (2005). Epigenetic Transgenerational Actions of Endocrine Disruptors and Male Fertility. *Science*, 308(5727):1466–1469.
- Ascierto, P. A., Kirkwood, J. M., Grob, J.-J., Simeone, E., Grimaldi, A. M., Maio, M., Palmieri, G., Testori, A., Marincola, F. M., and Mozzillo, N. (2012). The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine*, 10:85.
- Babinet, C., Farza, H., Morello, D., Hadchouel, M., and Pourcel, C. (1985). Specific expression of hepatitis B surface antigen (HBsAg) in transgenic mice. *Science*, 230(4730):1160–3.
- Bakiri, L. and Wagner, E. F. (2013). Mouse models for liver cancer. *Molecular Oncology*, 7(2):206–23.
- Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395.
- Bartolomei, M. S., Zemel, S., and Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. *Nature*, 351(6322):153–5.
- Bedford, M. T. and Clarke, S. G. (2009). Protein arginine methylation in mammals: who, what, and why. *Molecular Cell*, 33(1):1–13.
- Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–5.
- Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–96.
- Bender, S., Tang, Y., Lindroth, A. M., Hovestadt, V., Jones, D. T. W., Kool, M., Zapatka, M., Northcott, P. A., Sturm, D., Wang, W., Radlwimmer, B., Højfeldt, J. W., Truffaux, N., Castel, D., Schubert, S., Ryzhova, M., Seker-Cin, H., Gronych, J., Johann, P. D., Stark, S., Meyer, J., Milde, T., Schuhmann, M., Ebinger, M., Monoranu, C.-M., Ponnuswami, A., Chen, S., Jones, C., Witt, O., Collins, V. P., von Deimling, A., Jabado, N., Puget, S., Grill, J., Helin, K., Korshunov, A., Lichter, P., Monje, M., Plass, C., Cho, Y.-J., and Pfister, S. M. (2013). Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. *Cancer Cell*, 24(5):660–72.
- Berger, S. L., Kouzarides, T., Shiekhata, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, 23(7):781–783.
- Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- Bickmore, W. A. (2013). The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics*, 14(1):67–84.

- Blanco, N. D., Kruse, K., Erdmann, T., Staiger, A. M., Ott, G., Lenz, G., and Vaquerizas, J. M. (2018). Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *bioRxiv*, page 372789.
- Boj, S. F., Petrov, D., and Ferrer, J. (2010). Epistasis of Transcriptomes Reveals Synergism between Transcriptional Activators Hnf1 α and Hnf4 α . *PLoS Genetics*, 6(5):e1000970.
- Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156.
- Boyault, S., Rickman, D. S., de Reyniès, A., Balabaud, C., Rebouissou, S., Jeannot, E., Hérault, A., Saric, J., Belghiti, J., Franco, D., Bioulac-Sage, P., Laurent-Puig, P., and Zucman-Rossi, J. (2007). Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology*, 45(1):42–52.
- Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527.
- Broutier, L., Mastrogiovanni, G., Verstegen, M. M., Francies, H. E., Gavarró, L. M., Bradshaw, C. R., Allen, G. E., Arnes-Benito, R., Sidorova, O., Gaspersz, M. P., Georgakopoulos, N., Koo, B.-K., Dietmann, S., Davies, S. E., Praseedom, R. K., Lieshout, R., IJzermans, J. N. M., Wigmore, S. J., Saeb-Parsy, K., Garnett, M. J., van der Laan, L. J., and Huch, M. (2017). Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nature Medicine*, 23(12):1424–1435.
- Brown, G. (2015). GreyListChIP: Grey Lists – Mask Artefact Regions Based on ChIP Inputs. *R package version 1.6.0*.
- Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J., and Schwarz, M. (2008). Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutation Research*, 638(1-2):66–74.
- Burke, L. J., Hollemann, T., Pieler, T., and Renkawitz, R. (2002). Molecular cloning and expression of the chromatin insulator protein CTCF in *Xenopus laevis*. *Mechanisms of Development*, 113(1):95–8.
- C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396):2012–8.
- Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., Davidson, S., Edwards, M., Elvin, J. A., Hodel, K. P., Zahurancik, W. J., Suo, Z., Lipman, T., Wimmer, K., Kratz, C. P., Bowers, D. C., Laetsch, T. W., Dunn, G. P., Johanns, T. M., Grimmer, M. R., Smirnov, I. V., Larouche, V., Samuel, D., Bronsema, A., Osborn, M., Stearns, D., Raman, P., Cole, K. A., Storm, P. B., Yalon, M., Opocher, E., Mason, G., Thomas, G. A., Sabel, M., George, B., Ziegler, D. S.,

- Lindhorst, S., Issai, V. M., Constantini, S., Toledano, H., Elhasid, R., Farah, R., Dvir, R., Dirks, P., Huang, A., Galati, M. A., Chung, J., Ramaswamy, V., Irwin, M. S., Aronson, M., Durno, C., Taylor, M. D., Rechavi, G., Maris, J. M., Bouffet, E., Hawkins, C., Costello, J. F., Meyn, M. S., Pursell, Z. F., Malkin, D., Tabori, U., and Shlien, A. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. *Cell*, 171(5):1042–1056.
- Chambeyron, S. and Bickmore, W. A. (2004). Does looping and clustering in the nucleus regulate gene expression? *Current Opinion in Cell Biology*, 16(3):256–262.
- Chao, W., Huynh, K. D., Spencer, R. J., Davidow, L. S., and Lee, J. T. (2002). CTCF, a candidate trans-acting factor for X-inactivation choice. *Science*, 295(5553):345–7.
- Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472.
- Chen, H., Yu, H., Wang, J., Zhang, Z., Gao, Z., Chen, Z., Lu, Y., Liu, W., Jiang, D., Zheng, S. L., Wei, G.-H., Issacs, W. B., Feng, J., and Xu, J. (2015). Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci. *The Prostate*, 75(12):1264–76.
- Chien, A. J., Conrad, W. H., and Moon, R. T. (2009). A Wnt survival guide: from flies to human disease. *The Journal of Investigative Dermatology*, 129(7):1614–27.
- Chisari, F. V., Pinkert, C. A., Milich, D. R., Filippi, P., McLachlan, A., Palmiter, R. D., and Brinster, R. L. (1985). A transgenic mouse model of the chronic hepatitis B surface antigen carrier state. *Science*, 230(4730):1157–60.
- Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133.
- Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1):781.
- Connor, F., Rayner, T. F., Aitken, S. J., Feig, C., Lukk, M., Santoyo-Lopez, J., and Odom, D. T. (2018). Mutational landscape of a chemically-induced mouse model of liver cancer. *Journal of Hepatology*, 69(4):840–850.

- Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301.
- Cui, S., Hano, H., Sakata, A., Harada, T., Liu, T., Takai, S., and Ushigome, S. (1996). Enhanced CD34 expression of sinusoid-like vascular endothelial cells in hepatocellular carcinoma. *Pathology International*, 46(10):751–6.
- Dai, J., Zhu, M., Wang, C., Shen, W., Zhou, W., Sun, J., Liu, J., and Jin, G. (2015). Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population. *Scientific Reports*, 5:7833.
- Davison, A. C. A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.
- de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, 60(4):676–684.
- Dekker, J. and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. *Cell*, 164(6):1110–1121.
- Ding, J., Taylor, M. S., Jackson, A. P., and Reijns, M. A. M. (2015). Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq. *Nature Protocols*, 10(9):1433–1444.
- Dixon, J., Gorkin, D., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, 62(5):668–680.
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.
- Downen, J., Fan, Z., Hnisz, D., Ren, G., Abraham, B., Zhang, L., Weintraub, A., Schuijers, J., Lee, T., Zhao, K., and Young, R. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell*, 159(2):374–387.
- Duncan, A. W., Taylor, M. H., Hickey, R. D., Hanlon Newell, A. E., Lenzi, M. L., Olson, S. B., Finegold, M. J., and Grompe, M. (2010). The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature*, 467(7316):707–710.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.
- Egger, G., Liang, G., Aparicio, A., and Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–63.

- El-Serag, H. B. and Rudolph, K. L. (2007). Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. *Gastroenterology*, 132(7):2557–2576.
- Ernst, C., Odom, D. T., and Kutter, C. (2017). The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nature Communications*, 8(1):1411.
- Evan, G. I. and Littlewood, T. D. (1993). The role of c-myc in cell growth. *Current Opinion in Genetics & Development*, 3(1):44–9.
- Fedoriw, A. M., Stein, P., Svoboda, P., Schultz, R. M., and Bartolomei, M. S. (2004). Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science*, 303(5655):238–40.
- Fields, S. and Johnston, M. (2005). Cell biology. Whither model organism research? *Science*, 307(5717):1885–6.
- Filippova, G. N., Cheng, M. K., Moore, J. M., Truong, J.-P., Hu, Y. J., Nguyen, D. K., Tsuchiya, K. D., and Disteche, C. M. (2005). Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Developmental Cell*, 8(1):31–42.
- Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenko, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology*, 16(6):2802–13.
- Filippova, G. N., Lindblom, A., Meincke, L. J., Klenova, E. M., Neiman, P. E., Collins, S. J., Doggett, N. A., and Lobanenko, V. V. (1998). A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Genes, Chromosomes & Cancer*, 22(1):26–36.
- Fisher, D. I., Safrany, S. T., Strike, P., McLennan, A. G., and Cartwright, J. L. (2002). Nudix Hydrolases That Degrade Dinucleoside and Diphosphoinositol Polyphosphates Also Have 5-Phosphoribosyl 1-Pyrophosphate (PRPP) Pyrophosphatase Activity That Generates the Glycolytic Activator Ribose 1,5-Bisphosphate. *Journal of Biological Chemistry*, 277(49):47313–47317.
- Flavahan, W. A., Drier, Y., Liao, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., and Bernstein, B. E. (2015). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110–114.
- Flavahan, W. A., Gaskell, E., and Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science*, 357(6348):eaal2380.
- Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., and Campbell, P. J. (2017).

- COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783.
- Fraser, J., Williamson, I., Bickmore, W. A., and Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, 79(3):347–72.
- Fraser, P. (2006). Transcriptional control thrown for a loop. *Current Opinion in Genetics & Development*, 16(5):490–495.
- Frese, K. K. and Tuveson, D. A. (2007). Maximizing mouse cancer models. *Nature Reviews Cancer*, 7(9):654–658.
- Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, 49(12):1684–1692.
- Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., and Mirny, L. A. (2018). Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harbor Symposia on Quantitative Biology*, 82:45–55.
- Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., Gotoh, K., Ariizumi, S.-i., Wardell, C. P., Hayami, S., Nakamura, T., Aikata, H., Arihiro, K., Boroevich, K. A., Abe, T., Nakano, K., Maejima, K., Sasaki-Oku, A., Ohsawa, A., Shibuya, T., Nakamura, H., Hama, N., Hosoda, F., Arai, Y., Ohashi, S., Urushidate, T., Nagae, G., Yamamoto, S., Ueda, H., Tatsuno, K., Ojima, H., Hiraoka, N., Okusaka, T., Kubo, M., Marubashi, S., Yamada, T., Hirano, S., Yamamoto, M., Ohdan, H., Shimada, K., Ishikawa, O., Yamaue, H., Chayama, K., Miyano, S., Aburatani, H., Shibata, T., and Nakagawa, H. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nature Genetics*, 48(5):500–509.
- Furth, P. A., St Onge, L., Böger, H., Gruss, P., Gossen, M., Kistner, A., Bujard, H., and Hennighausen, L. (1994). Temporal control of gene expression in transgenic mice by a tetracycline-responsive promoter. *Proceedings of the National Academy of Sciences*, 91(20):9302–6.
- Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science*, 6384:102–105.
- Gentric, G. and Desdouets, C. (2014). Polyploidization in Liver Tissue. *The American Journal of Pathology*, 184(2):322–331.
- Ghirlando, R. and Felsenfeld, G. (2016). CTCF: making the right connections. *Genes & Development*, 30(8):881–891.
- Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3D genome. *Molecular Cell*, 49(5):773–82.

- Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M. T., Xie, L., Paulson, J. R., Earnshaw, W. C., Mirny, L. A., and Dekker, J. (2018). A pathway for mitotic chromosome formation. *Science*, 359(6376):eaao6135.
- Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–2.
- Guénet, J.-L. and Bonhomme, F. (2003). Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in Genetics*, 19(1):24–31.
- Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M. Q., Ren, B., Krainer, A. R., Maniatis, T., and Wu, Q. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4):900–910.
- Haag, T., Herkt, C. E., Walesch, S. K., Richter, A. M., and Dammann, R. H. (2014). The apoptosis associated tyrosine kinase gene is frequently hypermethylated in human cancer and is regulated by epigenetic mechanisms. *Genes & Cancer*, 5(9-10):365–74.
- Haarhuis, J. H., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., Krijger, P. H., Teunissen, H., Medema, R. H., van Steensel, B., Brummelkamp, T. R., de Wit, E., and Rowland, B. D. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, 169(4):693–707.
- Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D. A., Rozenblatt-Rosen, O., Zhang, F., and Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958.
- Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–3.
- Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.
- Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74.
- Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Biggs, D., Preece, C., Downes, D. J., Gosden, M., Sharpe, J. A., Sloane-Stanley, J. A., Hughes, J. R., Davies, B., and Higgs, D. R. (2017). Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature Cell Biology*, 19(8):952–961.
- Hardy, T. and Mann, D. A. (2016). Epigenetics in liver disease: from biology to therapeutics. *Gut*, 65(11):1895–1905.

- Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., Collins, V. P., and Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology*, 18(1):125.
- Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L. A., Sloane-Stanley, J. A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R. J., Smith, A. J. H., Wood, W. G., and Higgs, D. R. (2016). Genetic dissection of the α -globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903.
- Heath, H., Ribeiro De Almeida, C., Sleutels, F., Dingjan, G., Van De Nobelen, S., Jonkers, I., Ling, K.-W., Gribnau, J., Renkawitz, R., Grosveld, F., Hendriks, R. W., and Galjart, N. (2008). CTCF regulates cell cycle progression of $\alpha\beta$ T cells in the thymus. *The EMBO Journal*, 27(24):2839–2850.
- Heindryckx, F., Colle, I., and Van Vlierberghe, H. (2009). Experimental mouse models for hepatocellular carcinoma research. *International Journal of Experimental Pathology*, 90(4):367–86.
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–89.
- Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. (2016). Ensembl comparative genomics resources. *Database (Oxford)*, page bav096.
- Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., and Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Reports*, 2(2):345–57.
- Holliday, R. and Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–32.
- Hoogenraad, C. C., Koekkoek, B., Akhmanova, A., Krugers, H., Dortland, B., Miedema, M., van Alphen, A., Kistler, W. M., Jaegle, M., Koutsourakis, M., Van Camp, N., Verhoye, M., van der Linden, A., Kaverina, I., Grosveld, F., De Zeeuw, C. I., and Galjart, N. (2002). Targeted mutation of Cyln2 in the Williams syndrome critical region links CLIP-115 haploinsufficiency to neurodevelopmental abnormalities in mice. *Nature Genetics*, 32(1):116–127.

- Huch, M., Dorrell, C., Boj, S. F., van Es, J. H., Li, V. S. W., van de Wetering, M., Sato, T., Hamer, K., Sasaki, N., Finegold, M. J., Haft, A., Vries, R. G., Grompe, M., and Clevers, H. (2013). In vitro expansion of single Lgr5+ liver stem cells induced by Wnt-driven regeneration. *Nature*, 494(7436):247–250.
- Huch, M. and Koo, B.-K. (2015). Modeling mouse and human development using organoid cultures. *Development*, 142(18):3113–3125.
- Hudson, T. J., Anderson, W., Artz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Gutmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M. F., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O. M., Joly, Y., Kato, K., Kennedy, K. L., Nicolás, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clément, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., van de Vijver, M., Futreal, P. A., Aburatani, H., Bayés, M., Botwell, D. D. L., Campbell, P. J., Estivill, X., Grimmond, S. M., Gut, I., Hirst, M., López-Otín, C., Majumder, P., Marra, M., McPherson, J. D., Ning, Z., Puente, X. S., Ruan, Y., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Flicek, P., Getz, G., Guigó, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M., Li, Q., López-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B. F. F., Pearson, J. V., Quesada, V., Raphael, B. J., Sander, C., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Lathrop, M., Thomas, G., Yoshida, T., Axton, M., Gunter, C., Miller, L. J., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cros, A., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Chalmers, D. R. C., Hasel, K. W., Kaan, T. S. H., Lowrance, W. W., Masui, T., Rodriguez, L. L., Vergely, C., Bowtell, D. D. L., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. B., Gardiner, B. A., Kench, J. G., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., Gallinger, S., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., DePinho, R. A., Thayer, S., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R. E., Uhlén, M., Viksna, J., Ponten, F., Skryabin, K., Birney, E., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. A., Martin, S., Reis-Filho, J. S., Richardson, A. L., Sotiriou, C., Thoms, G., van't Veer, L., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Masson-Jacquemier, J. D., Pauporté, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Treilleux, I., Bioulac-Sage, P., Decaens, T., Franco, D., Gut, M., Samuel, D., Zucman-Rossi, J., Eils, R., Brors, B., Korbel, J. O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifemberger, G., Taylor, M. D., von Kalle, C., Majumder, P. P., Pederzoli, P., Lawlor, R. A., Delledonne, M., Bardelli, A., Gress, T., Klimstra, D., Zamboni, G., Nakamura, Y., Miyano, S., Fujimoto, A., Campo, E., de Sanjosé, S., Montserrat, E., González-Díaz, M., Jares, P., Himmelbauer, H., Himmelbaue, H.,

- Bea, S., Aparicio, S., Easton, D. F., Collins, F. S., Compton, C. C., Lander, E. S., Burke, W., Green, A. R., Hamilton, S. R., Kallioniemi, O. P., Ley, T. J., Liu, E. T., and Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, 464(7291):993–8.
- Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell*, 169(2):216–228.
- Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–56.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096):816–821.
- Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6):415–28.
- Jones, P. A. and Laird, P. W. (1999). Cancer epigenetics comes of age. *Nature Genetics*, 21(2):163–7.
- Kaiser, V. B., Taylor, M. S., Semple, C. A., Brenner, S., and Guilbaud, G. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genetics*, 12(8):e1006207.
- Kane, M. F., Loda, M., Gaida, G. M., Lipman, J., Mishra, R., Goldman, H., Jessup, J. M., and Kolodner, R. (1997). Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Research*, 57(5):808–811.
- Kang, J. S., Wanibuchi, H., Morimura, K., Gonzalez, F. J., and Fukushima, S. (2007). Role of CYP2E1 in Diethylnitrosamine-Induced Hepatocarcinogenesis In vivo. *Cancer Research*, 67(23):11141–11146.
- Kang, J. Y., Song, S. H., Yun, J., Jeon, M. S., Kim, H. P., Han, S. W., and Kim, T. Y. (2015). Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. *Oncogene*, 34(45):5677–84.
- Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.-P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J., and Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, 47(7):818–21.
- Kelleher, R. J., Flanagan, P. M., and Kornberg, R. D. (1990). A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell*, 61(7):1209–15.

- Kemp, C. J., Moore, J. M., Moser, R., Bernard, B., Teater, M., Smith, L. E., Rabaia, N. A., Gurley, K. E., Guinney, J., Busch, S. E., Shaknovich, R., Lobanenko, V. V., Liggitt, D., Shmulevich, I., Melnick, A., and Filippova, G. N. (2014). CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. *Cell Reports*, 7(4):1020–9.
- Kiselycznyk, C. and Holmes, A. (2011). All (C57BL/6) Mice are not Created Equal. *Frontiers in Neuroscience*, 5(10):1–3.
- Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D., and Kirschner, M. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.
- Klenova, E. M., Fagerlie, S., Filippova, G. N., Kretzner, L., Goodwin, G. H., Loring, G., Neiman, P. E., and Lobanenko, V. V. (1998). Characterization of the chicken CTCF genomic locus, and initial study of the cell cycle-regulated promoter of the gene. *The Journal of Biological Chemistry*, 273(41):26571–9.
- Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., Neiman, P. E., and Lobanenko, V. V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and Cellular Biology*, 13(12):7612–24.
- Koh, J. C., Loo, W. M., Goh, K. L., Sugano, K., Chan, W. K., Chiu, W. Y. P., Choi, M.-G., Gonlachanvit, S., Lee, W.-J., Lee, W. J. J., Lee, Y. Y., Lesmana, L. A., Li, Y.-M., Liu, C. J., Matsuura, B., Nakajima, A., Ng, E. K. W., Sollano, J. D., Wong, S. K. H., Wong, V. W. S., Yang, Y., Ho, K. Y., and Dan, Y. Y. (2016). Asian consensus on the relationship between obesity and gastrointestinal and liver diseases. *Journal of Gastroenterology and Hepatology*, 31(8):1405–1413.
- Kopp, F. and Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, 172(3):393–407.
- Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell*, 128(4):693–705.
- Lachner, M. and Jenuwein, T. (2002). The many faces of histone lysine methylation. *Current Opinion in Cell Biology*, 14(3):286–298.
- Lai, A. Y., Fatemi, M., Dhasarathy, A., Malone, C., Sobol, S. E., Geigerman, C., Jaye, D. L., Mav, D., Shah, R., Li, L., and Wade, P. A. (2010). DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas. *The Journal of Experimental Medicine*, 207(9):1939–50.
- Lallemand, Y., Luria, V., Haffner-Krausz, R., and Lonai, P. (1998). Maternally expressed PGK-Cre transgene as a tool for early and uniform activation of the Cre site-specific recombinase. *Transgenic Research*, 7(2):105–12.

- Lancaster, M. A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L. S., Hurles, M. E., Homfray, T., Penninger, J. M., Jackson, A. P., and Knoblich, J. A. (2013). Cerebral organoids model human brain development and microcephaly. *Nature*, 501(7467):373–379.
- Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.
- Lanier, M. H., Kim, T., and Cooper, J. A. (2015). CARMIL2 is a novel molecular connection between vimentin and actin essential for cell migration and invadopodia formation. *Molecular Biology of the Cell*, 26(25):4577–4588.
- Lappalainen, T. and Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics*, 18(7):441–451.
- Lathia, J. D., Mack, S. C., Mulkearns-Hubert, E. E., Valentim, C. L., and Rich, J. N. (2015). Cancer stem cells in glioblastoma. *Genes & Development*, 29(12):1203–1217.
- Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–2.
- Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.
- Lee, D., Tan, W., Anene, G., Li, P., Danh, T., Tiang, Z., Ng, S. L., Efthymios, M., Autio, M., Jiang, J., Fullwood, M., Prabhakar, S., and Foo, R. (2017). Gene neighbourhood integrity disrupted by CTCF loss in vivo. *bioRxiv*, page 187393.
- Leek, J., Johnson, W., Parker, H., Fertig, E., Jaffe, A., Storey, J., Zhang, Y., and Torres, L. (2017). sva: Surrogate Variable Analysis. *R package version 3.26.0*.
- Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., Bioulac-Sage, P., Prévôt, S., Azoulay, D., Paradis, V., Imbeaud, S., Deleuze, J.-F., and Zucman-Rossi, J. (2017). Mutational

- signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, 8(1):1315.
- Lewis, A. and Murrell, A. (2004). Genomic imprinting: CTCF protects the boundaries. *Current Biology*, 14(7):284–6.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Li, Y., Tang, Z.-Y., and Hou, J.-X. (2012). Hepatocellular carcinoma: insight from animal models. *Nature Reviews Gastroenterology & Hepatology*, 9(1):32–43.
- Liang, Y., Niederstrasser, H., Edwards, M., Jackson, C. E., and Cooper, J. A. (2009). Distinct Roles for CARMIL Isoforms in Cell Migration. *Molecular Biology of the Cell*, 20(24):5290–5305.
- Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93.
- Lilue, J., Doran, A. G., Fiddes, I. T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Czechanski, A., Danecek, P., Diekhans, M., Dolle, D.-D., Dunn, M., Durbin, R., Earl, D., Ferguson-Smith, A., Flicek, P., Flint, J., Frankish, A., Fu, B., Gerstein, M., Gilbert, J., Goodstadt, L., Harrow, J., Howe, K., Kolmogorov, M., Koenig, S., Lelliott, C., Loveland, J., Mott, R., Muir, P., Navarro, F., Odom, D., Park, N., Pelan, S., Phan, S. K., Quail, M., Reinholdt, L., Romoth, L., Shirley, L., Sisu, C., Sjoberg-Herrera, M., Stanke, M., Steward, C., Thomas, M., Threadgold, G., Thybert, D., Torrance, J., Wong, K., Wood, J., Yang, F., Adams, D. J., Paten, B., and Keane, T. M. (2018). Multiple laboratory mouse reference genomes define strain specific haplotypes and novel functional loci. *bioRxiv*, page 235838.
- Ling, J. Q., Li, T., Hu, J. F., Vu, T. H., Chen, H. L., Qiu, X. W., Cherry, A. M., and Hoffman, A. R. (2006). CTCF Mediates Interchromosomal Colocalization Between Igf2/H19 and Wsb1/Nf1. *Science*, 312(5771):269–272.
- Liu, G., Mattick, J. S., and Taft, R. J. (2013a). A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle*, 12(13):2061–72.

- Liu, L., De, S., and Michor, F. (2013b). DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature Communications*, 4:1502.
- Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550.
- Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., and Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12):1743–53.
- López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–217.
- Love, M. I., Anders, S., Kim, V., and Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, 4:1070.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.
- Lun, A. T. and Smyth, G. K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16(1):258.
- Lun, A. T. and Smyth, G. K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, 44(5):e45.
- Lun, A. T. L. and Smyth, G. K. (2014). De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Research*, 42(11):e95.
- Ma, S. and Rubin, B. P. (2014). Apoptosis-associated tyrosine kinase 1 inhibits growth and migration and promotes apoptosis in melanoma. *Laboratory Investigation*, 94(4):430–8.
- Macaulay, I. C., Teng, M. J., Haerty, W., Kumar, P., Ponting, C. P., and Voet, T. (2016). Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nature Protocols*, 11(11):2081–103.
- Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–7.
- Mack, S. C., Witt, H., Piro, R. M., Gu, L., Zuyderduyn, S., Stütz, A. M., Wang, X., Gallo, M., Garzia, L., Zayne, K., Zhang, X., Ramaswamy, V., Jäger, N., Jones, D. T. W., Sill, M., Pugh, T. J., Ryzhova, M., Wani, K. M., Shih, D. J. H., Head, R., Remke, M., Bailey, S. D., Zichner, T., Faria, C. C., Barszczyk, M., Stark, S., Seker-Cin, H., Hutter, S., Johann, P., Bender, S., Hovestadt, V., Tzaridis, T., Dubuc, A. M., Northcott, P. A., Peacock, J., Bertrand, K. C., Agnihotri, S., Cavalli, F. M. G., Clarke, I., Nethery-Brokk, K., Creasy, C. L., Verma, S. K., Koster, J., Wu, X., Yao,

- Y., Milde, T., Sin-Chan, P., Zuccaro, J., Lau, L., Pereira, S., Castelo-Branco, P., Hirst, M., Marra, M. A., Roberts, S. S., Fults, D., Massimi, L., Cho, Y. J., Van Meter, T., Grajkowska, W., Lach, B., Kulozik, A. E., von Deimling, A., Witt, O., Scherer, S. W., Fan, X., Muraszko, K. M., Kool, M., Pomeroy, S. L., Gupta, N., Phillips, J., Huang, A., Tabori, U., Hawkins, C., Malkin, D., Kongkham, P. N., Weiss, W. A., Jabado, N., Rutka, J. T., Bouffet, E., Korbel, J. O., Lupien, M., Aldape, K. D., Bader, G. D., Eils, R., Lichter, P., Dirks, P. B., Pfister, S. M., Korshunov, A., and Taylor, M. D. (2014). Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature*, 506(7489):445–450.
- Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A., and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.
- Maher, E. R. and Reik, W. (2000). Beckwith-Wiedemann syndrome: imprinting in clusters revisited. *Journal of Clinical Investigation*, 105(3):247–845.
- Mao, P., Brown, A. J., Malc, E. P., Mieczkowski, P. A., Smerdon, M. J., Roberts, S. A., and Wyrick, J. J. (2017). Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Research*, 27(10):1674–84.
- Margall-Ducos, G., Celton-Morizur, S., Couton, D., Bregerie, O., and Desdouets, C. (2007). Liver tetraploidization is controlled by a new process of incomplete cytokinesis. *Journal of Cell Science*, 120(20):3633–3639.
- Marinov, G. K., Kundaje, A., Park, P. J., and Wold, B. J. (2014). Large-Scale Quality Analysis of Published ChIP-seq Data. *G3: Genes, Genomes, Genetics*, 4(2):209–223.
- Maronpot, R. R. (2009). Biological Basis of Differential Susceptibility to Hepatocarcinogenesis among Mouse Strains. *Journal of Toxicologic Pathology*, 22(1):11–33.
- Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.
- Maximin, S., Ganeshan, D. M., Shanbhogue, A. K., Dighe, M. K., Yeh, M. M., Kolokythas, O., Bhargava, P., and Lalwani, N. (2014). Current update on combined hepatocellular-cholangiocarcinoma. *European Journal of Radiology Open*, 1:40–48.
- McAlister, G. C., Nusinow, D. P., Jedrychowski, M. P., Wühr, M., Huttlin, E. L., Erickson, B. K., Rad, R., Haas, W., and Gygi, S. P. (2014). MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Analytical Chemistry*, 86(14):7150–7158.
- McGrath, J. and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1):179–83.

- Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigo, R. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665.
- Melton, C., Reuter, J. A., Spacek, D. V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, 47(7):710–716.
- Messerschmidt, D. M., Knowles, B. B., and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes & Development*, 28(8):812–28.
- Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Grealley, J. M., Gut, I., Houseman, E. A., Izzi, B., Kelsey, K. T., Meissner, A., Milosavljevic, A., Siegmund, K. D., Bock, C., and Irizarry, R. A. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–955.
- Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S. T., Munhall, A., Grewe, B., Bartkuhn, M., Arnold, R., Burke, L. J., Renkawitz-Pohl, R., Ohlsson, R., Zhou, J., Renkawitz, R., and Lobanenko, V. (2005). CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Reports*, 6(2):165–70.
- Moore, B. L., Aitken, S., and Semple, C. A. (2015). Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biology*, 16(1):110.
- Moore, J. M., Rabaia, N. A., Smith, L. E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C. M., Collins, S. J., Kemp, C. J., Lobanenko, V. V., and Filippova, G. N. (2012). Loss of maternal CTCF is associated with peri-implantation lethality of *Ctcf* null embryos. *PloS One*, 7(4):e34915.
- Murphy, D. A. and Courtneidge, S. A. (2011). The 'ins' and 'outs' of podosomes and invadopodia: characteristics, formation and function. *Nature Reviews Molecular Cell Biology*, 12(7):413–426.
- Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67.
- Nagano, T., Lubling, Y., Yaffe, E., Wingett, S. W., Dean, W., Tanay, A., and Fraser, P. (2015). Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nature Protocols*, 10(12):1986–2003.
- Nam, S. W., Park, J. Y., Ramasamy, A., Shevade, S., Islam, A., Long, P. M., Park, C. K., Park, S. E., Kim, S. Y., Lee, S. H., Park, W. S., Yoo, N. J., Liu, E. T., Miller, L. D., and Lee, J. Y. (2005). Molecular changes from dysplastic nodule to hepatocellular carcinoma through gene expression profiling. *Hepatology*, 42(4):809–818.

- Nanney, D. L. (1958). Epigenetic Control Systems. *Proceedings of the National Academy of Sciences*, 44(7):712–7.
- Nassar, D., Latil, M., Boeckx, B., Lambrechts, D., and Blanpain, C. (2015). Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nature Medicine*, 21(8):946–54.
- Naugler, W. E., Sakurai, T., Kim, S., Maeda, S., Kim, K., Elsharkawy, A. M., and Karin, M. (2007). Gender Disparity in Liver Cancer Due to Sex Differences in MyD88-Dependent IL-6 Production. *Science*, 317(5834):121–124.
- Newell, P., Villanueva, A., Friedman, S. L., Koike, K., and Llovet, J. M. (2008). Experimental models of hepatocellular carcinoma. *Journal of Hepatology*, 48(5):858–879.
- Nicholls, R. D., Saitoh, S., and Horsthemke, B. (1998). Imprinting in Prader-Willi and Angelman syndromes. *Trends in Genetics*, 14(5):194–200.
- Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerod, A., Aparicio, S. A. J. R., Tutt, A., Sieuwerts, A. M., Borg, A., Thomas, G., Salomon, A. V., Richardson, A. L., Borresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., and Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.
- Nora, E. P., Goloborodko, A., Valton, A.-L., Dekker, J., Mirny, L. A., Bruneau Correspondence, B. G., Gibcus, J. H., Uebersohn, A., Abdennur, N., and Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169:930–944.
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706.
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., Macisaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-Specific Transcriptional Regulation has Diverged Significantly between Human and Mouse. *Nature Genetics*, 39(6):730–732.
- Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, 17(9):520–7.
- Oinonen, T. and Lindros, K. O. (1998). Zonation of hepatic cytochrome P-450 expression and regulation. *Biochemical Journal*, 329(1):17–35.

- Oki, M., Aihara, H., and Ito, T. (2007). Role of histone phosphorylation in chromatin dynamics and its implications in diseases. *Sub-cellular Biochemistry*, 41:319–36.
- Owenab, E. H., Loguea, S. F., Rasmussena, D. L., and Wehnerac, J. M. (1997). Assessment of learning by the Morris water task and fear conditioning in inbred mouse strains and F1 hybrids: implications of genetic background for single gene mutations and quantitative trait loci analyses. *Neuroscience*, 80(4):1087–99.
- Owens, G. L., Todesco, M., Drummond, E. B. M., Yeaman, S., and Rieseberg, L. H. (2018). A novel post hoc method for detecting index switching finds no evidence for increased switching on the Illumina HiSeq X. *Molecular Ecology Resources*, 18(1):169–175.
- Palstra, R.-J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. (2003). The β -globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, 35(2):190–194.
- Pant, V., Kurukuti, S., Pugacheva, E., Shamsuddin, S., Mariano, P., Renkawitz, R., Klenova, E., Lobanenko, V., and Ohlsson, R. (2004). Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. *Molecular and Cellular Biology*, 24(8):3497–504.
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290.
- Park, C., Qian, W., and Zhang, J. (2012). Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Reports*, 13(12):1123–9.
- Parthun, M. R. (2007). Hat1: the emerging cellular roles of a type B histone acetyltransferase. *Oncogene*, 26(37):5319–28.
- Pegg, A. E. (1990). Mammalian O6-alkylguanine-DNA alkyltransferase: regulation and importance in response to alkylating carcinogenic and therapeutic agents. *Cancer Research*, 50(19):6119–29.
- Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., and Wong, J. W. H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*, 532(7598):259–263.
- Peschansky, V. J. and Wahlestedt, C. (2014). Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics*, 9(1):3–12.
- Petukhova, G. V., Pezza, R. J., Vanevski, F., Ploquin, M., Masson, J.-Y., and Camerini-Otero, R. D. (2005). The Hop2 and Mnd1 proteins act in concert with Rad51 and Dmc1 in meiotic recombination. *Nature Structural & Molecular Biology*, 12(5):449–453.
- Phillips, J. E. and Corces, V. G. (2009). CTCF: Master Weaver of the Genome. *Cell*, 137(7):1194–1211.

- Phillips-Cremins, J., Sauria, M., Sanyal, A., Gerasimova, T., Lajoie, B., Bell, J., Ong, C.-T., Hookway, T., Guo, C., Sun, Y., Bland, M., Wagstaff, W., Dalton, S., McDevitt, T., Sen, R., Dekker, J., Taylor, J., and Corces, V. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, 153(6):1281–1295.
- Poetsch, A. R., Boulton, S. J., and Luscombe, N. (2018). Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *bioRxiv*, page 168153.
- Postic, C., Shiota, M., Niswender, K. D., Jetton, T. L., Chen, Y., Moates, J. M., Shelton, K. D., Lindner, J., Cherrington, A. D., and Magnuson, M. A. (1999). Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *The Journal of Biological Chemistry*, 274(1):305–15.
- Poulos, R. C., Thoms, J. A., Guan, Y. F., Unnikrishnan, A., Pimanda, J. E., and Wong, J. W. (2016). Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Reports*, 17(11):2865–2872.
- Prawitt, D., Enklaar, T., Gärtner-Rupprecht, B., Spangenberg, C., Oswald, M., Lausch, E., Schmidtke, P., Reutzel, D., Fees, S., Lucito, R., Korzon, M., Brozek, I., Limon, J., Housman, D. E., Pelletier, J., and Zabel, B. (2005). Microdeletion of target sites for insulator protein CTCF in a chromosome 11p15 imprinting center in Beckwith-Wiedemann syndrome and Wilms' tumor. *Proceedings of the National Academy of Sciences*, 102(11):4085–90.
- Puccini, J., Dorstyn, L., and Kumar, S. (2013). Genetic background and tumour susceptibility in mouse models. *Cell Death & Differentiation*, 20(7):964–964.
- Pugacheva, E. M., Kwon, Y.-W., Hukriede, N. A., Pack, S., Flanagan, P. T., Ahn, J.-C., Park, J. A., Choi, K.-S., Kim, K.-W., Loukinov, D., Dawid, I. B., and Lobanenko, V. V. (2006). Cloning and characterization of zebrafish CTCF: Developmental expression patterns, regulation of the promoter region, and evolutionary aspects of gene organization. *Gene*, 375:26–36.
- Qi, Y., Chen, X., Chan, C.-y., Li, D., Yuan, C., Yu, F., Lin, M. C., Yew, D. T., Kung, H.-F., and Lai, L. (2008). Two-dimensional differential gel electrophoresis/analysis of diethylnitrosamine induced rat hepatocellular carcinoma. *International Journal of Cancer*, 122(12):2682–2688.
- Qin, F., Song, Z., Babiceanu, M., Song, Y., Facemire, L., Singh, R., Adli, M., and Li, H. (2015). Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in human prostate cells. *PLoS Genetics*, 11(2):e1005001.
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2.
- Rabbitts, T. H. (1994). Chromosomal translocations in human cancer. *Nature*, 372(6502):143–9.

- Rabbitts, T. H. (2009). Commonality but diversity in cancer gene fusions. *Cell*, 137(3):391–5.
- Rad, R., Cadiñanos, J., Rad, L., Varela, I., Strong, A., Kriegel, L., Constantino-Casas, F., Eser, S., Hieber, M., Seidler, B., Price, S., Fraga, M. F., Calvanese, V., Hoffman, G., Ponstingl, H., Schneider, G., Yusa, K., Grove, C., Schmid, R. M., Wang, W., Vassiliou, G., Kirchner, T., McDermott, U., Liu, P., Saur, D., and Bradley, A. (2013). A genetic progression model of Braf(V600E)-induced intestinal tumorigenesis reveals targets for therapeutic intervention. *Cancer Cell*, 24(1):15–29.
- Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.
- Rajewsky, M. F., Dauber, W., and Frankenberg, H. (1966). Liver carcinogenesis by diethylnitrosamine in the rat. *Science*, 152(3718):83–5.
- Rao, S. S., Huang, S.-C., Hilaire, B. G. S., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., Casellas, R., Lander, E. S., and Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2):305–320.
- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680.
- Recillas-Targa, F., Rosa-Velázquez, I. A., Soto-Reyes, E., and Benítez-Bribiesca, L. (2006). Epigenetic boundaries of tumour suppressor gene promoters: the CTCF connection and its role in carcinogenesis. *Journal of Cellular and Molecular Medicine*, 10(3):554–568.
- Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32.
- Reményi, A., Schöler, H. R., and Wilmanns, M. (2004). Combinatorial control of gene expression. *Nature Structural & Molecular Biology*, 11(9):812–5.
- Ren, G., Jin, W., Cui, K., Rodriguez, J., Hu, G., Zhang, Z., Larson, D. R., and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*, 67(6):1049–1058.
- Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M., Bergen, I., Thongjuea, S., Lenhard, B., van IJcken, W., Grosveld, F., Galjart, N., Soler, E., and Hendriks, R. (2011). The DNA-Binding Protein CTCF Limits Proximal V κ Recombination and Restricts κ Enhancer Interactions to the Immunoglobulin κ Light Chain Locus. *Immunity*, 35(4):501–513.

- Riggs, A. and Porter, T. (1996). Overview of epigenetic mechanisms. In Russo VEA, Martienssen R, and Riggs AD, editors, *Epigenetic Mechanisms of Gene Regulation*, pages 29–45. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry*, 81:145–66.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.
- Roeder, R. G. and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224(5216):234–7.
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31.
- Rubio-Perez, C., Tamborero, D., Schroeder, M., Antolín, A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2015). In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, 27(3):382–396.
- Ruiz-Velasco, M. and Zaugg, J. B. (2017). Structure meets function: How chromatin organisation conveys functionality. *Current Opinion in Systems Biology*, 1:129–136.
- Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, 532(7598):264–267.
- Sabarinathan, R., Pich, O., Martincorena, I., Rubio-Perez, C., Juul, M., Wala, J., Schumacher, S., Shapira, O., Sidiropoulos, N., Waszak, S., Tamborero, D., Mularoni, L., Rheinbay, E., Hornshøj, H., Deu-Pons, J., Muinos, F., Bertl, J., Guo, Q., Weischenfeldt, J., Korbel, J. O., Getz, G., Campbell, P. J., Pedersen, J. S., Beroukhi, R., Perez-Gonzalez, A., Lopez-Bigas, N., Group, P. D., Interpretation, F., and Net, I. P.-C. A. o. W. G. (2017). The whole-genome panorama of cancer drivers. *bioRxiv*, page 190330.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–25.
- Saldaña-Meyer, R., González-Buendía, E., Guerrero, G., Narendra, V., Bonasio, R., Recillas-Targa, F., and Reinberg, D. (2014). CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes & Development*, 28(7):723–34.
- Saldaña-Meyer, R. and Recillas-Targa, F. (2011). Transcriptional and epigenetic regulation of the p53 tumor suppressor gene. *Epigenetics*, 6(9):1068–77.

- Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):6456–65.
- Sandoval, J. and Esteller, M. (2012). Cancer epigenomics: beyond genomics. *Current Opinion in Genetics & Development*, 22(1):50–55.
- Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C. T., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature*, 419(6905):407–411.
- Sauer, B. and Henderson, N. (1988). Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. *Proceedings of the National Academy of Sciences*, 85(14):5166–70.
- Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817.
- Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., Flicek, P., and Odom, D. T. (2010a). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research*, 20(5):578–88.
- Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1-2):335–48.
- Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010b). Five vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040.
- Schneider, C., Teufel, A., Yevsa, T., Staib, F., Hohmeyer, A., Walenda, G., Zimmermann, H. W., Vucur, M., Huss, S., Gassler, N., Wasmuth, H. E., Lira, S. A., Zender, L., Luedde, T., Trautwein, C., and Tacke, F. (2012). Adaptive immunity suppresses formation and progression of diethylnitrosamine-induced liver cancer. *Gut*, 61(12):1733–43.
- Schoenfelder, S., Clay, I., and Fraser, P. (2010). The transcriptional interactome: gene expression in 3D. *Current Opinion in Genetics & Development*, 20(2):127–33.
- Scholzen, T. and Gerdes, J. (2000). The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology*, 182(3):311–322.
- Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., Calatayud, A.-L., Pinyol, R., Pelletier, L., Balabaud, C., Laurent, A., Blanc, J.-F., Mazzaferro, V., Calvo,

- F., Villanueva, A., Nault, J.-C., Bioulac-Sage, P., Stratton, M. R., Llovet, J. M., and Zucman-Rossi, J. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature Genetics*, 47(5):505–511.
- Schulze, K., Nault, J.-C., and Villanueva, A. (2016). Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *Journal of Hepatology*, 65(5):1031–1042.
- Schuster-Böckler, B. and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–7.
- Schwalie, P. C., Ward, M. C., Cain, C. E., Faure, A. J., Gilad, Y., Odom, D. T., and Flicek, P. (2013). Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biology*, 14(12):R148.
- Schwartz-Arad, D., Zajicek, G., and Bartfeld, E. (1989). The streaming liver IV: DNA content of the hepatocyte increases with its age. *Liver*, 9(2):93–99.
- Schwartzman, O. and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726.
- Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51.
- Sharma, S., Kelly, T. K., and Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36.
- Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–86.
- Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature Genetics*, 38(11):1348–1354.
- Skinner, M. K. (2014). Endocrine disruptor induction of epigenetic transgenerational inheritance of disease. *Molecular and Cellular Endocrinology*, 398(1-2):4–12.
- Smith, G. S., Walford, R. L., and Mickey, M. R. (1973). Lifespan and Incidence of Cancer and Other Diseases in Selected Long-Lived Inbred Mice and Their F1 Hybrids². *Journal of the National Cancer Institute*, 50(5):1195–1213.
- Smith, K. S., Liu, L. L., Ganesan, S., Michor, F., and De, S. (2017). Nuclear topology modulates the mutational landscapes of cancer genomes. *Nature Structural & Molecular Biology*, 24(11):1000–1006.

- Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S. M., Schroth, G. P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal*, 32(24):3119–3129.
- Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development*, 20(17):2349–54.
- Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393–5.
- Stark, R. and Brown, G. (2011). DiffBind: differential binding analysis of ChIP-Seq peak data. *R package version 2.2.7*.
- Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D. J., Talianidis, I., Marioni, J. C., Flicek, P., and Odom, D. T. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–40.
- Stricker, S. H., Köferle, A., and Beck, S. (2016). From profiles to function in epigenomics. *Nature Reviews Genetics*, 18(1):51–66.
- Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. *Genes & Development*, 12(5):599–606.
- Sugino, T., Yamaguchi, T., Hoshi, N., Kusakabe, T., Ogura, G., Goodison, S., and Suzuki, T. (2008). Sinusoidal tumor angiogenesis is a key component in hepatocellular carcinoma metastasis. *Clinical & Experimental Metastasis*, 25(7):835–41.
- Surani, M. A., Barton, S. C., and Norris, M. L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308(5959):548–50.
- Taby, R. and Issa, J.-P. J. (2010). Cancer epigenetics. *CA: A Cancer Journal for Clinicians*, 60(6):376–92.
- Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, 99(6):3740–5.
- Taketo, M., Schroeder, A. C., Mobraaten, L. E., Gunning, K. B., Hanten, G., Fox, R. R., Roderick, T. H., Stewart, C. L., Lilly, F., Hansen, C. T., and Al., E. (1991). FVB/N: an inbred mouse strain preferable for transgenic analyses. *Proceedings of the National Academy of Sciences*, 88(6):2065–9.
- Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, 12(4):e1004873.

- Terakawa, T., Bisht, S., Eeftens, J. M., Dekker, C., Haering, C. H., and Greene, E. C. (2017). The condensin complex is a mechanochemical motor that translocates along DNA. *Science*, 358(6363):672–676.
- The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.
- The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.
- Thoolen, B., Maronpot, R. R., Harada, T., Nyska, A., Rousseaux, C., Nolte, T., Malarkey, D. E., Kaufmann, W., Küttler, K., Deschl, U., Nakae, D., Gregson, R., Vinlove, M. P., Brix, A. E., Singh, B., Belpoggi, F., and Ward, J. M. (2010). Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicologic pathology*, 38(7 Suppl):5S–81S.
- Thybert, D., Roller, M., Navarro, F. C. P., Fiddes, I., Streeter, I., Feig, C., Martin-Galvez, D., Kolmogorov, M., Janoušek, V., Akanni, W., Aken, B., Aldridge, S., Chakrapani, V., Chow, W., Clarke, L., Cummins, C., Doran, A., Dunn, M., Goodstadt, L., Howe, K., Howell, M., Josselin, A.-A., Karn, R. C., Laukaitis, C. M., Jingtao, L., Martin, F., Muffato, M., Nachtweide, S., Quail, M. A., Sisu, C., Stanke, M., Stefflova, K., Van Oosterhout, C., Veyrunes, F., Ward, B., Yang, F., Yazdanifar, G., Zadissa, A., Adams, D. J., Brazma, A., Gerstein, M., Paten, B., Pham, S., Keane, T. M., Odom, D. T., and Flicek, P. (2018). Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Research*, 28(4):448–459.
- Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.
- Tsai, C.-L., Rowntree, R. K., Cohen, D. E., and Lee, J. T. (2008). Higher order chromatin structure at the X-inactivation center via looping DNA. *Developmental Biology*, 319(2):416–425.
- Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigartyo, C. A.-k., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H. P.-h., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.
- Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., and Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, 357(6352):eaan2507.

- Varki, A. and Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Research*, 15(12):1746–58.
- Velázquez-Hernández, N., Reyes-Romero, M. A., Barragán-Hernández, M., Guerrero-Romero, F., Rodríguez-Moran, M., Aguilar-Durán, M., and Lazalde Medina, B. (2015). BORIS and CTCF are overexpressed in squamous intraepithelial lesions and cervical cancer. *Genetics and Molecular Research*, 14(2):6094–6100.
- Ventura, A. and Jacks, T. (2009). MicroRNAs and cancer: short RNAs go a long way. *Cell*, 136(4):586–91.
- Verna, L., Whysner, J., and Williams, G. M. (1996). N-nitrosodiethylamine mechanistic data and risk assessment: bioactivation, DNA-adduct formation, mutagenicity, and tumor initiation. *Pharmacology & Therapeutics*, 71(1-2):57–81.
- Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, 10(8):1297–1309.
- Vietri Rudan, M. and Hadjur, S. (2015). Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. *Trends in Genetics*, 31(11):651–660.
- Villar, D., Berthelot, C., Aldridge, S., Rayner, T., Lukk, M., Pignatelli, M., Park, T., Deaville, R., Erichsen, J., Jasinska, A., Turner, J., Bertelsen, M., Murchison, E., Flicek, P., and Odom, D. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566.
- Villar, D., Flicek, P., and Odom, D. T. (2014). Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Reviews Genetics*, 15(4):221–33.
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–58.
- Waddington, C. H. (1942). The epigenotype. *Endeavour*, 1:18–20.
- Wan, L.-B., Pan, H., Hannenhalli, S., Cheng, Y., Ma, J., Fedoriw, A., Lobanenko, V., Latham, K. E., Schultz, R. M., and Bartolomei, M. S. (2008). Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development*, 135(16):2729–2738.
- Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R. S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. *Nature*, 431(7010):873–878.
- Wang, K. C. and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular Cell*, 43(6):904–14.
- Wang, Y. and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Molecular Cell*, 58(4):598–609.

- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42:1001–6.
- Westcott, P. M. K., Halliwill, K. D., To, M. D., Rashid, M., Rust, A. G., Keane, T. M., Delrosario, R., Jen, K.-Y., Gurley, K. E., Kemp, C. J., Fredlund, E., Quigley, D. A., Adams, D. J., and Balmain, A. (2014). The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature*, 517(7535):489–492.
- White, B. D., Chien, A. J., and Dawson, D. W. (2012). Dysregulation of Wnt/ β -catenin signaling in gastrointestinal cancers. *Gastroenterology*, 142(2):219–32.
- Wilkinson, L. S., Davies, W., and Isles, A. R. (2007). Genomic imprinting effects on brain development and function. *Nature Reviews Neuroscience*, 8(11):832–43.
- Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, 4:1310.
- Wu, A. R., Wang, J., Streets, A. M., and Huang, Y. (2017). Single-Cell Transcriptional Analysis. *Annual Review of Analytical Chemistry*, 10(1):439–462.
- Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews Genetics*, 12(8):542–53.
- Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Molecular and Cellular Biology*, 31(11):2174–83.
- Xu, N., Donohoe, M. E., Silva, S. S., and Lee, J. T. (2007). Evidence that homologous X-chromosome pairing requires transcription and Ctf protein. *Nature Genetics*, 39(11):1390–6.
- Yang, F., Deng, X., Ma, W., Berletch, J. B., Rabaia, N., Wei, G., Moore, J. M., Filippova, G. N., Xu, J., Liu, Y., Noble, W. S., Shendure, J., and Disteche, C. M. (2015). The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biology*, 16(1):52.
- Yang, H., Wang, H., Shivalila, C., Cheng, A., Shi, L., and Jaenisch, R. (2013). One-Step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR/Cas-Mediated Genome Engineering. *Cell*, 154(6):1370–1379.
- Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J.,

- Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, 44(D1):710–6.
- Yoon, Y. S., Jeong, S., Rong, Q., Park, K.-Y., Chung, J. H., and Pfeifer, K. (2007). Analysis of the H19ICR insulator. *Molecular and Cellular Biology*, 27(9):3499–510.
- Yoshida, Y., Adachi, E., Fukiya, K., Iwai, K., and Tanaka, K. (2005). Glycoprotein-specific ubiquitin ligases recognize N-glycans in unfolded substrates. *EMBO Reports*, 6(3):239–244.
- Yost, C., Torres, M., Miller, J. R., Huang, E., Kimelman, D., and Moon, R. T. (1996). The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in *Xenopus* embryos by glycogen synthase kinase 3. *Genes & Development*, 10(12):1443–54.
- You, J. S. and Jones, P. A. (2012). Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell*, 22(1):9–20.
- Yu, N. Y.-L., Hallström, B. M., Fagerberg, L., Ponten, F., Kawaji, H., Carninci, P., Forrest, A., FANTOM Consortium, T., Hayashizaki, Y., Uhlén, M., and Daub, C. O. (2015). Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Research*, 43(14):6787–6798.
- Zeng, Q.-A., Qiu, J., Zou, R., Li, Y., Li, S., Li, B., Huang, P., Hong, J., Zheng, Y., Lao, X., and Yuan, Y. (2012). Clinical features and outcome of multiple primary malignancies involving hepatocellular carcinoma: a long-term follow-up study. *BMC Cancer*, 12:148.
- Zhang, S., Zhou, K., Luo, X., Li, L., Tu, H.-C., Sehgal, A., Nguyen, L. H., Zhang, Y., Gopal, P., Tarlow, B. D., Siegwart, D. J., and Zhu, H. (2018). The Polyploid State Plays a Tumor-Suppressive Role in the Liver. *Developmental Cell*, 44(4):447–459.
- Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. *EMBO Reports*, 16(11):1467–81.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., Liu, X. S., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., Peters, J., Croce, L. D., Wutz, A., Hendrich, B., Klenerman, D., and Laue, E. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.
- Zhang, Y.-W., Brognard, J., Coughlin, C., You, Z., Dolled-Filhart, M., Aslanian, A., Manning, G., Abraham, R. T., and Hunter, T. (2009). The F Box Protein Fbx6 Regulates Chk1 Stability and Cellular Sensitivity to Replication Stress. *Molecular Cell*, 35(4):442–453.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643.

- Zucman-Rossi, J., Villanueva, A., Nault, J.-C., and Llovet, J. M. (2015). Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. *Gastroenterology*, 149(5):1226–1239.
- Zuin, J., Dixon, J. R., van der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. A., van IJcken, W. F. J., Grosveld, F. G., Ren, B., and Wendt, K. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences*, 111(3):996–1001.

List of differentially expressed genes

Table 1 Differential expression results of RNA-seq experiments performed in wild-type and *Ctcf* hemizygous mouse embryonic fibroblasts. Differential expression analysis of the RNA-seq, performed using DESeq2, controlling for both the known and hidden batch effects. baseMean is the mean of normalised counts of all samples, normalising for sequencing depth. Genes were considered significantly differentially expressed if their adjusted p-value was lower than 0.05.

Ensembl gene ID	Gene	Chr	Start	End	baseMean	Log ₂ FoldChange	p value	Adjusted p value
ENSMUSG00000044080	S100a1	3	90511034	90514392	327.7706364	-0.834560741	5.13E-18	8.28E-14
ENSMUSG00000028031	Dkk2	3	132085292	132180304	1405.339917	-0.696904508	2.55E-16	2.05E-12
ENSMUSG00000062380	Tubb3	8	123411424	123422015	1656.794211	-0.557823603	6.24E-16	3.35E-12
ENSMUSG00000052353	Cemip	7	83932857	84086502	13392.54319	0.547217942	2.38E-15	8.94E-12
ENSMUSG00000057751	Megf6	4	154170730	154275713	635.4414121	0.769519438	2.77E-15	8.94E-12
ENSMUSG00000027009	Itga4	2	79255426	79333123	1324.029672	-0.645207921	6.01E-14	1.62E-10
ENSMUSG00000027827	Kcnab1	3	65109384	65378223	349.7014279	-0.711456325	8.38E-14	1.93E-10
ENSMUSG00000054690	Emcn	3	137341067	137432185	156.5416287	-0.784392219	3.50E-13	7.05E-10
ENSMUSG00000027347	Rasgrp1	2	117279993	117343001	93.63284528	0.761366767	1.14E-12	2.03E-09
ENSMUSG00000034161	Scx	15	76457438	76459468	694.5414233	0.547242439	2.30E-11	3.37E-08
ENSMUSG00000045827	Serpinb9	13	33003250	33017957	1883.135215	0.487883922	2.17E-11	3.37E-08
ENSMUSG00000070327	Rnf213	11	119393100	119487418	5424.823794	-0.456292951	4.57E-11	6.14E-08
ENSMUSG00000076441	Ass1	2	31470207	31520672	4678.467071	-0.49656652	1.22E-10	1.51E-07
ENSMUSG00000031825	Crispld2	8	119992438	120052793	1393.667197	-0.532977475	1.46E-10	1.56E-07
ENSMUSG00000056758	Hmga2	10	120361275	120476469	4116.823937	-0.460716584	1.54E-10	1.56E-07
ENSMUSG00000074896	Ifit3	19	34583531	34588731	272.0103622	-0.643617453	1.36E-10	1.56E-07
ENSMUSG0000005698	Ctcf	8	105636568	105682922	1579.87131	-0.457815754	2.54E-10	2.40E-07
ENSMUSG00000029671	Wnt16	6	22288227	22298522	905.201436	0.60125982	4.58E-10	4.11E-07
ENSMUSG0000000120	Ngfr	11	95568818	95587735	197.1525474	0.658935595	4.92E-10	4.18E-07
ENSMUSG00000015766	Eps8	6	137477245	137654876	2278.150851	-0.44639976	5.84E-10	4.71E-07
ENSMUSG00000072621	Sfn10-ps	11	83028130	83040042	67.2856275	-0.656404501	6.32E-10	4.85E-07
ENSMUSG00000061353	Cxcl12	6	117168535	117181367	20110.88015	-0.412378116	6.98E-10	5.11E-07
ENSMUSG00000027087	Itgav	2	83724397	83806916	18570.59646	-0.351619782	3.80E-09	2.67E-06
ENSMUSG00000029778	Adcyap1r1	6	55451978	55501451	233.0226462	-0.623787885	5.49E-09	3.69E-06
ENSMUSG00000050357	Rltpr	8	105690906	105698178	26.7341123	0.514268484	5.79E-09	3.74E-06
ENSMUSG00000035385	Ccl2	11	82035571	82037453	779.2040822	-0.485277037	6.46E-09	4.01E-06
ENSMUSG00000026109	Trmeff2	1	50900647	51187270	561.4977155	-0.560674028	7.40E-09	4.42E-06
ENSMUSG00000050587	Cd44	2	102811141	102901665	12408.64082	-0.43986125	1.03E-08	5.96E-06
ENSMUSG00000025150	Cbr2	11	120729489	120732114	1706.469054	0.463213545	1.35E-08	7.50E-06
ENSMUSG00000036412	Arsi	18	60911780	60918561	532.2748015	0.562510283	1.47E-08	7.88E-06
ENSMUSG00000021614	Vcan	13	89655312	89742509	21654.92819	-0.360126007	2.16E-08	1.12E-05
ENSMUSG00000032359	Ctsh	9	90054152	90076089	1185.159355	-0.403522197	3.50E-08	1.76E-05
ENSMUSG00000029061	Mmp23	4	155650655	155653384	1964.753004	-0.440869707	4.41E-08	2.16E-05
ENSMUSG00000022544	Eef2kmt	16	5244152	5255983	506.1441695	0.463625316	5.51E-08	2.61E-05
ENSMUSG00000032492	Pth1r	9	110722085	110747145	595.0992815	0.469161569	6.27E-08	2.89E-05
ENSMUSG00000047414	Flrt2	12	95692226	95785215	7013.008857	-0.353549686	9.53E-08	4.15E-05
ENSMUSG00000054580	Pla2r1	2	60417543	60553308	658.6523683	-0.511934523	9.53E-08	4.15E-05
ENSMUSG00000028370	Pappa	4	65124174	65357509	17361.8346	0.39254509	1.02E-07	4.33E-05
ENSMUSG00000024747	Aldh1a7	19	20692953	20727562	139.1952329	-0.566649659	1.22E-07	5.03E-05
ENSMUSG00000030036	Mogs	6	83115496	83118898	2231.283832	0.367100288	1.46E-07	5.90E-05
ENSMUSG00000022114	Spry2	14	105891949	105896819	888.0415273	-0.415840026	1.60E-07	6.20E-05
ENSMUSG00000045777	Ifitm10	7	142325837	142373753	87.15619384	0.563975709	1.62E-07	6.20E-05
ENSMUSG00000035356	Nfkbiz	16	55811375	55838899	1400.031916	-0.392742378	1.86E-07	6.97E-05

Differentially expressed genes (cont.) 1								
Ensembl gene ID	Gene	Chr	Start	End	Mean	Log ₂ FoldChange	p value	Adjusted p value
ENSMUSG00000022464	Slc38a4	15	96994823	97055956	2612.144702	-0.40650444	1.94E-07	7.10E-05
ENSMUSG00000029664	Tfpi2	6	3962595	3988919	229.5366558	-0.549442076	2.35E-07	8.42E-05
ENSMUSG00000022799	Arhgap31	16	38598340	38713274	1387.061476	-0.348778658	2.52E-07	8.64E-05
ENSMUSG000000103585	Pcdhgb4	18	37720369	37841870	516.7890821	0.481967035	2.51E-07	8.64E-05
ENSMUSG00000005357	Slc1a6	10	78780496	78814825	1497.542603	0.421615633	2.92E-07	9.80E-05
ENSMUSG00000005397	Nid1	13	13437602	13512275	22540.12252	-0.296782038	3.10E-07	0.000102139
ENSMUSG000000001493	Meox1	11	101877510	101894374	500.8871802	0.449004298	3.31E-07	0.000106861
ENSMUSG00000016200	Syt14	1	192891233	193035775	195.3019047	0.529038541	3.45E-07	0.000108624
ENSMUSG000000030616	Syt12	7	90302252	90410719	1803.593124	-0.358656625	3.50E-07	0.000108624
ENSMUSG000000025492	Ifitm3	7	141009586	141010770	4180.106245	-0.353287459	3.81E-07	0.000115967
ENSMUSG00000049420	Tmem200a	10	25991186	26079052	340.3231611	-0.526988482	5.07E-07	0.000151445
ENSMUSG000000026547	Tagln2	1	172500047	172507380	4865.751659	-0.315757003	5.51E-07	0.000161662
ENSMUSG00000041559	Fmod	1	134037254	134048277	1192.804191	0.452635099	5.76E-07	0.00016574
ENSMUSG000000001123	Lgals9	11	78962974	78984946	512.9420761	-0.461757975	5.94E-07	0.000168149
ENSMUSG000000025504	Eps8l2	7	141338880	141363020	191.2155767	-0.536002828	6.18E-07	0.000171878
ENSMUSG00000041596	Nlrp5-ps	7	14530652	14622479	234.997219	0.497723962	6.53E-07	0.000178366
ENSMUSG000000021253	Tgfb3	12	86056744	86079041	13582.02791	0.307180553	8.92E-07	0.000239822
ENSMUSG00000053846	Lipg	18	74939322	74961263	547.0001425	-0.443530541	9.50E-07	0.000251122
ENSMUSG000000031391	L1cam	X	73853778	73896105	334.8813357	0.481549112	1.10E-06	0.000287039
ENSMUSG000000043289	Mei4	9	81863670	82206007	131.7684453	0.516838024	1.21E-06	0.000308707
ENSMUSG00000049723	Mmp12	9	7344381	7369499	249.1602682	-0.521844884	1.23E-06	0.000308707
ENSMUSG000000000275	Trim25	11	88999376	89020293	1019.466578	-0.375065653	1.51E-06	0.000373984
ENSMUSG000000029648	Flt1	5	147561604	147726011	1892.713187	-0.406941254	1.74E-06	0.000425758
ENSMUSG000000022122	Ednrb	14	103814625	103844173	296.8078163	-0.498954804	1.85E-06	0.00044443
ENSMUSG000000043943	Naalad2	9	18321951	18402995	83.69413571	-0.510860992	1.90E-06	0.00045023
ENSMUSG00000041658	Rragb	X	153139981	153171943	122.157914	0.510510938	2.14E-06	0.000500772
ENSMUSG000000015312	Gadd45b	10	80930091	80932204	1323.091768	0.386973458	2.21E-06	0.000510017
ENSMUSG000000029121	Crmp1	5	37241940	37292133	204.8444027	-0.4876146	2.35E-06	0.000533153
ENSMUSG000000020027	Socs2	10	95385362	95417180	279.8906471	-0.482858098	2.51E-06	0.00056184
ENSMUSG000000006731	B4galnt1	10	127165156	127172340	799.0366279	-0.380441249	2.55E-06	0.000562757
ENSMUSG000000029231	Pdgfra	5	75152292	75198215	2193.1792	-0.38361821	2.59E-06	0.000563832
ENSMUSG000000032420	Nt5e	9	88327197	88372092	581.4099365	0.412851207	2.81E-06	0.000603844
ENSMUSG000000052727	Map1b	13	99421464	99516602	8155.249681	-0.295171936	3.05E-06	0.000647242
ENSMUSG00000013846	St3gal1	15	67102875	67113992	468.5961735	-0.422557612	3.27E-06	0.000684297
ENSMUSG000000024084	Qpct	17	79051906	79090243	288.8812849	-0.470397535	3.31E-06	0.000684905
ENSMUSG000000021678	F2r1	13	95511732	95525240	366.799389	-0.450088191	3.62E-06	0.000739082
ENSMUSG000000031557	Plekha2	8	25039144	25102194	1343.402313	-0.340137939	3.70E-06	0.000745632
ENSMUSG000000064345	mt-Nd2	MT	3914	4951	2810.039625	-0.390517493	4.11E-06	0.000818001
ENSMUSG000000030827	Fgf21	7	45613907	45615490	117.5041029	0.491233162	5.71E-06	0.001122483
ENSMUSG000000020023	Tmcc3	10	94311949	94590956	555.9607442	-0.38699573	5.86E-06	0.001124608
ENSMUSG000000036109	Mbnl3	X	51117269	51206532	491.3732573	-0.441227241	5.81E-06	0.001124608
ENSMUSG000000030468	Siglecg	7	43408204	43418358	519.2261293	-0.426030315	6.17E-06	0.001170965
ENSMUSG00000003066	Gas7	11	67455437	67688990	674.4684719	-0.411850634	6.62E-06	0.001241988
ENSMUSG000000038932	Tcf15	2	180621956	180642708	415.4583111	0.464458327	6.88E-06	0.001275052
ENSMUSG000000026399	Cd55	1	130439027	130462744	649.2070936	-0.454598263	7.02E-06	0.001287221
ENSMUSG000000006360	Crip1	12	113146316	113153879	412.4452981	-0.424873209	7.24E-06	0.001311143
ENSMUSG000000033544	Angptl1	1	156838562	156861078	215.2232128	-0.475929158	7.77E-06	0.001391977
ENSMUSG000000034765	Dusp5	19	53529109	53542431	772.0216032	-0.39676695	7.98E-06	0.001414061
ENSMUSG000000025921	Rdh10	1	16105774	16133734	974.7566534	0.342008029	8.61E-06	0.001509081
ENSMUSG000000026042	Col5a2	1	45374321	45503282	237346.7503	0.291872784	8.94E-06	0.001549884
ENSMUSG000000009376	Met	6	17463800	17573980	1113.996245	-0.361361632	9.07E-06	0.001555097
ENSMUSG000000032368	Zic1	9	91358058	91365810	365.4673467	0.462237146	9.59E-06	0.001596907
ENSMUSG000000053062	Jam2	16	84774123	84825928	211.1981342	-0.464812362	9.42E-06	0.001596907
ENSMUSG000000074415	2610203C20Rik	9	41376397	41617772	2336.49848	-0.28548231	9.61E-06	0.001596907
ENSMUSG0000000054720	Lrrc8c	5	105519388	105613018	1377.563495	-0.355356371	9.90E-06	0.001628926
ENSMUSG000000031616	Ednra	8	77663031	77724464	331.797627	-0.468009521	1.01E-05	0.00165194
ENSMUSG000000002325	Irf9	14	55603571	55610030	938.0007915	-0.35391027	1.22E-05	0.00194943
ENSMUSG000000047735	Samd9l	6	3372257	3399572	1652.134436	-0.371535207	1.22E-05	0.00194943
ENSMUSG000000036986	Pml	9	58218076	58249786	802.8430406	-0.341189777	1.26E-05	0.001993477
ENSMUSG000000020357	Flt4	11	49609263	49652739	277.160189	-0.434896052	1.37E-05	0.002143034
ENSMUSG000000022512	Cldn1	16	26356642	26371841	258.5499926	-0.448021568	1.39E-05	0.002150849
ENSMUSG000000062151	Unc13c	9	73479422	73968966	73.32136551	-0.443369845	1.47E-05	0.002255065
ENSMUSG000000051951	Xkr4	1	3205901	3671498	91.1478604	0.465914579	1.52E-05	0.002288072
ENSMUSG000000057534	Gm15698	11	88964658	88966931	130.5730439	-0.453348603	1.50E-05	0.002288072
ENSMUSG000000062098	Btbd3	2	138256565	138589292	1295.635035	-0.309077769	1.71E-05	0.002546172
ENSMUSG000000038418	Egr1	18	34859823	34864984	464.3403941	-0.417154391	1.77E-05	0.002615656
ENSMUSG000000059970	Hspa2	12	76404176	76406934	855.4904407	0.34591884	2.10E-05	0.003081735
ENSMUSG000000096965	330005D01Rik	17	5798657	5803240	183.2336572	-0.453036761	2.24E-05	0.003247111
ENSMUSG000000024501	Dpysl3	18	43320979	43438286	20372.5474	-0.279250597	2.31E-05	0.00332163
ENSMUSG000000025068	Gsto1	19	47854970	47864790	9105.01178	-0.261065929	2.45E-05	0.003471258
ENSMUSG000000087247	Fam150a	1	6359218	6394731	763.1317967	0.36529284	2.45E-05	0.003471258

Differentially expressed genes (cont.) 1

Ensembl gene ID	Gene	Chr	Start	End	Mean	Log ₂ FoldChange	p value	Adjusted p value
ENSMUSG00000057596	Trim30d	7	104470014	104507849	213.3824073	-0.441369707	2.54E-05	0.003568188
ENSMUSG00000060988	Galnt13	2	54436317	55118309	311.5604182	-0.431538593	2.63E-05	0.003655547
ENSMUSG00000030208	Emp1	6	135362545	135383173	17406.7732	-0.243845519	2.66E-05	0.003669473
ENSMUSG00000033063	Cntnap3	13	64737591	64903888	28.39363177	-0.371652855	2.72E-05	0.003722291
ENSMUSG00000028217	Cdh17	4	11758147	11817895	201.7429846	0.437587176	3.09E-05	0.004180726
ENSMUSG00000020601	Trib2	12	15791727	15816877	1453.241475	-0.300597937	3.18E-05	0.004278242
ENSMUSG00000040010	Slc7a5	8	121881150	121907694	4005.822434	0.329676208	3.23E-05	0.004299276
ENSMUSG0000001657	Hoxc8	15	102990607	102994111	492.5635641	-0.406368737	3.49E-05	0.004566025
ENSMUSG00000028885	Smpd13b	4	132732966	132757252	93.46094942	-0.446775174	3.46E-05	0.004566025
ENSMUSG00000036009	Mettl25	10	105763189	105841380	263.2763806	0.409098976	3.51E-05	0.004566025
ENSMUSG00000027533	Fabp5	3	10012548	10016607	2511.94828	0.268226708	3.64E-05	0.004689502
ENSMUSG00000040263	Klhdc4	8	121796313	121829569	840.9006279	-0.321971195	3.88E-05	0.004961366
ENSMUSG00000025743	Sdc3	4	130792537	130826319	3652.770721	-0.257145434	4.59E-05	0.005823934
ENSMUSG00000046711	Hmga1	17	27556620	27563674	1750.873358	-0.34246055	4.94E-05	0.006219496
ENSMUSG00000061436	Hipk2	6	38694390	38876165	665.9314876	-0.332289651	5.18E-05	0.00648051
ENSMUSG00000014599	Csf1	3	107741048	107760469	6373.066601	-0.232331652	6.13E-05	0.007508878
ENSMUSG00000019960	Dusp6	10	99263231	99267488	1458.047239	-0.326373965	6.11E-05	0.007508878
ENSMUSG00000027784	Ppm1l	3	69316861	69560802	1226.133439	-0.312585403	6.15E-05	0.007508878
ENSMUSG00000024063	Lbh	17	72918305	72941942	5968.800506	-0.272353716	6.22E-05	0.00754376
ENSMUSG00000018417	Myo1b	1	51749765	51916071	3252.137917	-0.275719851	6.41E-05	0.007660451
ENSMUSG00000028020	Glr3	3	80843599	80913660	517.4559555	0.350691936	6.41E-05	0.007660451
ENSMUSG00000041731	Pgm5	19	24683016	24861855	1560.220208	-0.29522535	6.52E-05	0.00772979
ENSMUSG00000045868	Gvin1	7	106156556	106215326	150.3620541	-0.429998848	6.66E-05	0.007837066
ENSMUSG000000031402	Mpp1	X	75109733	75131016	3314.569525	-0.268792302	6.75E-05	0.007883692
ENSMUSG00000020689	Itgb3	11	104608000	104670476	1835.907719	0.303754982	6.88E-05	0.007980378
ENSMUSG00000021136	Smoc1	12	81026808	81186414	841.914684	-0.331993122	7.05E-05	0.008115114
ENSMUSG00000068566	Myadm	7	3289080	3300442	6560.488938	-0.269199501	7.29E-05	0.008332777
ENSMUSG00000007888	Crif1	8	70493158	70504081	2377.07306	0.27690881	7.40E-05	0.008339329
ENSMUSG00000030220	Arhgd1b	6	136923655	136941899	2558.09283	-0.346019869	7.38E-05	0.008339329
ENSMUSG000000039959	Hip1	5	135406531	135545120	1776.904422	-0.267606779	7.73E-05	0.008659126
ENSMUSG00000047281	Sfn	4	133600556	133602168	234.8185726	0.403430665	7.79E-05	0.008659126
ENSMUSG000000031626	Sorbs2	8	45507788	45827906	893.4432762	-0.371446086	8.12E-05	0.008964671
ENSMUSG00000027907	S100a11	3	93520488	93526287	1319.236656	-0.290411744	8.27E-05	0.009067171
ENSMUSG00000034751	Mast4	13	102732486	103334497	1499.619075	-0.287294782	8.35E-05	0.009092464
ENSMUSG00000026888	Grb14	2	64912476	65024987	1268.542177	-0.340881473	8.51E-05	0.009204501
ENSMUSG00000059493	Nhs	X	161833296	162159730	734.9067199	-0.360515757	8.98E-05	0.009648686
ENSMUSG000000039953	C1stn1	4	149586468	149648899	3021.383186	0.240765896	9.25E-05	0.009877881
ENSMUSG00000022665	Ccdc80	16	45093402	45127924	31870.09798	0.262016907	9.51E-05	0.010086494
ENSMUSG00000006205	Htra1	7	130936111	130985660	11635.7386	0.310798347	9.97E-05	0.010504556
ENSMUSG00000047216	Cdh19	1	110888326	110977584	1037.78863	0.419109547	0.000100422	0.010514971
ENSMUSG00000004655	Aqp1	6	55336432	55348555	1026.026556	-0.360718472	0.000102744	0.010688651
ENSMUSG00000026482	Rgl1	1	152516760	152766351	1739.430927	-0.291312114	0.000105462	0.010831697
ENSMUSG000000032899	Styk1	6	131299142	131353597	739.0761148	-0.370668564	0.00010538	0.010831697
ENSMUSG00000073418	C4b	17	34728380	34743882	295.7249142	0.415841583	0.000106753	0.010894861
ENSMUSG00000030249	Abcc9	6	142587862	142702315	100.8621008	-0.417078809	0.000108542	0.01100783
ENSMUSG00000040274	Cdk6	5	3341485	3531008	2303.302049	-0.29497398	0.000109611	0.011046692
ENSMUSG00000040033	Stat2	10	128270576	128292849	886.9793467	-0.339267923	0.000112133	0.01123069
ENSMUSG00000027188	Pamr1	2	102550012	102643041	673.8500103	0.357283484	0.00011416	0.011363198
ENSMUSG00000035245	Eogt	6	97110024	97149182	1237.435728	-0.27063652	0.000115141	0.011390455
ENSMUSG00000028179	Cth	3	157894248	157925077	1382.400807	0.315337014	0.000116205	0.01142561
ENSMUSG00000002257	Def6	17	28207778	28228608	362.8016127	-0.371431311	0.000118449	0.011575747
ENSMUSG000000041073	Nacad	11	6597823	6606053	195.0115986	-0.401982975	0.000119436	0.011601802
ENSMUSG000000106106	Rn18s-rs5	17	39846353	39848827	768.5767177	-0.336579553	0.000120649	0.0116495
ENSMUSG00000078606	Gm4070	7	105895139	105953967	107.6709642	-0.413170185	0.000124578	0.011957292
ENSMUSG000000048503	Tmem136	9	43108653	43116570	107.9589146	-0.411627267	0.000126652	0.012084444
ENSMUSG000000021319	Sfrp4	13	19623175	19632821	226.6674115	0.40658048	0.000128062	0.012100689
ENSMUSG00000040957	Cables1	18	11839220	11945627	161.5661405	0.404403477	0.000128324	0.012100689
ENSMUSG00000079481	Nhs12	X	101849385	102092055	213.5507486	-0.393969876	0.000133797	0.01254345
ENSMUSG00000029322	Plac8	5	100553725	100572245	668.0818713	-0.315656023	0.000134913	0.012575002
ENSMUSG00000033676	Gabrb3	7	57419692	57828802	188.7221149	0.391536604	0.000136172	0.012619394
ENSMUSG00000026051	1500015O10Rik	1	43730602	43742578	1105.319963	-0.330290134	0.000137779	0.012695389
ENSMUSG00000028019	Pdgfr	3	81036416	81214040	2188.305776	0.267165672	0.000140971	0.012909162
ENSMUSG00000028607	Cpt2	4	107903981	107923610	1271.275856	0.290472155	0.000141784	0.012909162
ENSMUSG00000047143	Dmrta2	4	109978053	109983687	40.29223962	0.386907185	0.000143302	0.012909162
ENSMUSG00000047747	Rnf150	8	82863356	83091271	2408.867219	-0.244716183	0.000143271	0.012909162
ENSMUSG0000000031722	Hp	8	109575130	109579172	48.48509068	-0.379262085	0.000144483	0.012943307
ENSMUSG00000033306	Lpp	16	24393350	24992576	26133.48462	-0.258091395	0.000150963	0.013449061
ENSMUSG00000029636	Waf3	5	146384985	146473615	191.0510972	-0.402308398	0.000152482	0.013509767
ENSMUSG00000028037	Ifi44	3	151730922	151749960	129.4130241	-0.40663702	0.000154545	0.01361766
ENSMUSG00000027656	Wisp2	2	163820861	163833146	7796.061746	0.387647517	0.000159766	0.014001253
ENSMUSG00000021250	Fos	12	85473890	85477273	340.9705728	-0.386409887	0.000161057	0.014038067

Differentially expressed genes (cont.) 1								
Ensembl gene ID	Gene	Chr	Start	End	Mean	Log ₂ FoldChange	p value	Adjusted p value
ENSMUSG000000005125	Ndrp1	15	66929321	66969640	1135.068111	0.269804131	0.000167494	0.014366181
ENSMUSG000000040170	Fmo2	1	162874317	162898726	100.8261271	-0.388305903	0.000166695	0.014366181
ENSMUSG000000062488	Ifit3b	19	34607970	34613401	75.5191183	-0.399052809	0.000166459	0.014366181
ENSMUSG000000067219	Nipal1	5	72647795	72671078	543.5785269	-0.34525465	0.000172063	0.014680021
ENSMUSG000000017724	Etv4	11	101769742	101785371	414.1297662	-0.362580522	0.000176194	0.014953309
ENSMUSG000000030921	Trim30a	7	104409025	104465193	221.6307452	-0.402761637	0.000179156	0.014968316
ENSMUSG000000041959	S100a10	3	93555080	93564643	4636.464742	-0.227480355	0.000178169	0.014968316
ENSMUSG000000046982	Tshz1	18	84011627	84086404	1513.986668	-0.25060169	0.000178768	0.014968316
ENSMUSG000000020431	Adcy1	11	7063489	7178506	595.6419584	-0.333864684	0.000182784	0.015192721
ENSMUSG000000021403	Serpinb9b	13	33027416	33041884	658.096708	-0.362498345	0.000183734	0.015193381
ENSMUSG000000049502	Dtx3l	16	35926511	35939151	509.0817086	-0.328275341	0.000185473	0.015258953
ENSMUSG000000004098	Col5a3	9	20770050	20815067	2672.743647	0.317517761	0.000191231	0.015573792
ENSMUSG000000009418	Nav1	1	135434580	135607295	562.4667028	-0.327180148	0.000191232	0.015573792
ENSMUSG0000000045573	Penk	4	4133531	4138819	2865.544352	-0.39015596	0.000193127	0.015649147
ENSMUSG000000044948	Cfap43	19	47737561	47919299	1844.978907	-0.277466512	0.000196729	0.015801259
ENSMUSG000000075707	Dio3	12	110279068	110281097	235.1630199	0.39132347	0.000196965	0.015801259
ENSMUSG000000006800	Sulf2	2	166073089	166155663	5671.017392	-0.292277661	0.000202443	0.016080716
ENSMUSG000000037736	Limch1	5	66745827	67057158	1936.418908	0.286264548	0.000202144	0.016080716
ENSMUSG0000000033207	Mamdc2	19	23302609	23448322	559.6013832	-0.353274159	0.000210227	0.016617231
ENSMUSG000000061666	Gdpc1	11	87033867	87074062	501.5603093	-0.331197813	0.000211643	0.016647539
ENSMUSG000000027397	Slc20a1	2	129198764	129211616	8154.937877	0.252068947	0.000215104	0.016837628
ENSMUSG000000023034	Nr4a1	15	101266846	101274792	466.3078692	-0.366276412	0.000220283	0.017058734
ENSMUSG000000054293	A630033H20Rik	X	107148927	107173661	98.64921243	-0.39755261	0.000220298	0.017058734
ENSMUSG000000060126	Tpt1	14	75845093	75848525	39058.33849	0.211140365	0.000221102	0.017058734
ENSMUSG0000000107317	Gm19719	5	149282752	149287868	23.26159872	-0.29642032	0.000227919	0.017500932
ENSMUSG000000027035	Cers6	2	68861441	69114282	4696.031381	-0.229094776	0.000246682	0.018501118
ENSMUSG0000000030107	Usp18	6	121245906	121270917	109.6395469	-0.387941536	0.000245807	0.018501118
ENSMUSG000000045083	Lingo2	4	35706647	36951747	214.3763946	-0.38085041	0.000244083	0.018501118
ENSMUSG0000000054793	Cadm4	7	24482023	24504539	119.971927	0.393946404	0.000244119	0.018501118
ENSMUSG0000000063430	Wscd2	5	113490333	113589725	202.6731494	0.394413674	0.000244766	0.018501118
ENSMUSG000000031226	Pbdc1	X	105079756	105117090	1253.824343	0.309226833	0.000256874	0.019087976
ENSMUSG000000074129	Rpl13a	7	45125565	45128745	36584.53348	0.225590862	0.000256864	0.019087976
ENSMUSG0000000031465	Angpt2	8	18690263	18741562	462.6833559	-0.348615136	0.000258963	0.019154975
ENSMUSG000000028864	Hgf	5	16553495	16620152	670.1868223	-0.367364639	0.000270375	0.019907723
ENSMUSG000000021892	Sh3bp5	14	31359880	31436078	504.5237297	-0.344394371	0.000272267	0.019955946
ENSMUSG000000020354	Sgcd	11	46896253	47988969	654.465766	-0.31656248	0.000281698	0.020553791
ENSMUSG000000021127	Zfp361l	12	80107760	80113013	608.7608515	-0.315317327	0.000290394	0.02108427
ENSMUSG000000029207	Apbb2	5	66298703	66618784	3784.515656	-0.225176423	0.000291584	0.02108427
ENSMUSG000000031997	Trpc6	9	8544196	8680565	54.53390216	-0.344947417	0.000295951	0.021209837
ENSMUSG0000000055401	Fbxo6	4	148145716	148152140	1129.768982	0.256456968	0.000295621	0.021209837
ENSMUSG00000103472	Pcdhga7	18	37714764	37841873	406.077515	-0.326996018	0.000297289	0.021211472
ENSMUSG000000065254	Gm23973	7	103271550	103271870	5471.865313	0.246147798	0.000300189	0.02132398
ENSMUSG0000000006627	Sema4f	6	82911885	82939769	50.66612402	0.3784419	0.00030524	0.021587675
ENSMUSG000000019966	Kitl	10	100015630	100100413	348.4464909	-0.373991093	0.000318786	0.022447246
ENSMUSG0000000036782	Klhl13	X	23219271	23365082	1667.795823	0.297838304	0.000324919	0.022779664
ENSMUSG000000030717	Nupr1	7	126623249	126630861	4925.928684	0.244622118	0.000334191	0.023328234
ENSMUSG0000000019577	Pdk4	6	5483351	5496309	286.4853213	-0.358799211	0.000357224	0.024828599
ENSMUSG000000025746	Il6	5	30013114	30019981	75.51299278	-0.377899212	0.000365908	0.025323017
ENSMUSG000000037108	Zcwpw1	5	137787798	137822621	154.2482613	0.384004726	0.000375045	0.025844417
ENSMUSG0000000030556	Lrrc28	7	67513410	67645268	566.2425385	-0.304985829	0.000379296	0.025915894
ENSMUSG000000062393	Dgkk	X	6779306	6948363	105.0975901	-0.382783466	0.000378692	0.025915894
ENSMUSG0000000054408	Spcs3	8	54520433	54529998	7081.579715	0.211690994	0.000382514	0.02602549
ENSMUSG0000000031530	Dcn	10	97479500	97518162	2353.860352	-0.280209572	0.000388048	0.026291067
ENSMUSG000000028024	Enpep	3	129269175	129332720	424.7652309	-0.358205203	0.000394701	0.026408962
ENSMUSG0000000031530	Dusp4	8	34807297	34819894	661.0339332	-0.32031298	0.000396402	0.026408962
ENSMUSG000000036676	Tmtc3	10	100443902	100487350	5078.946528	-0.211809851	0.000392766	0.026408962
ENSMUSG000000036377	C530008M17Rik	5	76656512	76873554	298.5866625	-0.342879251	0.000396755	0.026436693
ENSMUSG000000029826	Zc3hav1	6	38305286	38354603	1765.286199	-0.279198105	0.000401421	0.026637517
ENSMUSG000000030122	Ptms	6	124913681	124920103	3770.488549	-0.251282452	0.000405504	0.026798136
ENSMUSG0000000033355	Rtp4	16	23609919	23614222	91.19031189	-0.37496975	0.000454589	0.029919356
ENSMUSG000000053332	Gas5	1	161034422	161038539	7592.96487	0.217857139	0.000458394	0.029925521
ENSMUSG000000064341	mt-Nd1	MT	2751	3707	29504.22525	-0.267637993	0.000458203	0.029925521
ENSMUSG0000000025762	Larp1b	3	40950354	41040234	545.1439593	0.299469047	0.00047806	0.031083541
ENSMUSG000000027171	Prrg4	2	104830741	104849876	709.5575315	-0.330503951	0.000483412	0.031305322
ENSMUSG0000000034997	Htr2a	14	74640840	74706859	1061.065463	-0.326788287	0.000486212	0.031360684
ENSMUSG0000000062991	Nrg1	8	31814551	32884029	1390.620165	0.263869992	0.000491594	0.031581463
ENSMUSG000000020303	Stc2	11	31357307	31370074	439.9714033	0.336512435	0.000498861	0.031921158
ENSMUSG000000006356	Crip2	12	113140236	113145506	1576.092848	-0.249519198	0.000523925	0.033392244
ENSMUSG000000021097	Climn	12	104763114	104865076	258.9523211	0.367255633	0.000529379	0.033607218
ENSMUSG0000000019124	Scrrn1	6	54501173	54566489	1487.323047	0.264034434	0.000537263	0.033974207
ENSMUSG000000039316	Rftn1	17	49992257	50190674	927.2332749	-0.276026003	0.000541041	0.034079263

Differentially expressed genes (cont.) 1

Ensembl gene ID	Gene	Chr	Start	End	Mean	Log ₂ FoldChange	p value	Adjusted p value
ENSMUSG00000060969	Irx1	13	71957921	71963723	512.4032745	-0.332451569	0.000558348	0.035032539
ENSMUSG00000019806	Aig1	10	13647054	13868980	1149.511141	-0.26941207	0.000563084	0.035077697
ENSMUSG00000053647	Gper1	5	139423151	139427800	79.88044029	0.372511101	0.000563419	0.035077697
ENSMUSG000000087365	C430049B03Rik	X	53055046	53057190	269.5852266	-0.347805465	0.00058231	0.036114449
ENSMUSG000000062661	Ncs1	2	31245823	31295989	2752.064036	-0.214218978	0.000585405	0.036167274
ENSMUSG000000031647	Mfap3l	8	60632827	60676729	911.8002989	-0.262516424	0.000596903	0.036734199
ENSMUSG000000068748	Ptprz1	6	22875502	23052916	1838.692869	-0.25025218	0.000599138	0.036734199
ENSMUSG00000024277	Mapre2	18	23752333	23893861	717.810252	-0.277857049	0.000602773	0.03681712
ENSMUSG00000022090	Pdlim2	14	70164218	70177681	1049.619005	0.257475482	0.000610602	0.037014851
ENSMUSG000000037379	Spon2	5	33198184	33218455	198.8100285	-0.365218175	0.000609366	0.037014851
ENSMUSG000000030753	Prkrir	7	98703103	98718062	3182.277293	-0.205519561	0.000613656	0.037060667
ENSMUSG000000032312	Csk	9	57626647	57645653	1124.687261	-0.253324087	0.000617353	0.037144824
ENSMUSG000000072941	Sod3	5	52363791	52371418	420.4708011	-0.325404636	0.000632423	0.0377697
ENSMUSG000000106918	Mrpl33	5	31596935	31664384	1007.449065	0.265764804	0.000631026	0.0377697
ENSMUSG000000039518	Cdsn	17	35552128	35557180	296.9254502	0.340306029	0.000653783	0.038901281
ENSMUSG00000001349	Cnn1	9	22099281	22109630	9779.731118	0.256343923	0.000665894	0.039476242
ENSMUSG000000006219	Fblim1	4	141576062	141606096	1229.794684	-0.238294406	0.000669353	0.039535984
ENSMUSG000000000325	Arvcf	16	18348182	18407076	709.5390214	0.299410874	0.000684708	0.040148799
ENSMUSG000000096847	Tmem151b	17	45541940	45549677	19.1100961	-0.278773845	0.000682506	0.040148799
ENSMUSG000000046186	Cd109	9	78615546	78716253	6633.658596	-0.242482252	0.00070662	0.041283532
ENSMUSG000000028413	B4galnt1	4	40804602	40854005	3545.733926	-0.229663146	0.000711797	0.041435842
ENSMUSG000000018076	Med13l	5	118560679	118765438	1880.746289	-0.22972795	0.000720363	0.041783668
ENSMUSG000000048647	Exd1	2	119516505	119547627	89.05552336	0.360636001	0.000727186	0.042028224
ENSMUSG000000025507	Pidd1	7	141438113	141444025	265.323458	-0.340155614	0.000735849	0.042377007
ENSMUSG000000069874	Irgm2	11	58199618	58222782	255.6965618	-0.347119378	0.000749829	0.043028458
ENSMUSG000000015647	Lama5	2	180176373	180225859	3896.174807	0.258885131	0.000776663	0.044253342
ENSMUSG000000031207	Msn	X	96096042	96168552	19390.6378	-0.197519891	0.000774862	0.044253342
ENSMUSG000000038508	Gdf15	8	70629393	70632456	185.9543568	0.344033608	0.000780349	0.044306803
ENSMUSG000000067274	Rplp0	5	115559467	115563727	63270.88582	0.196634836	0.000783728	0.044342494
ENSMUSG000000035493	Tgfb1	13	56609603	56639339	3473.796683	0.235395323	0.000814325	0.045912577
ENSMUSG000000003032	Klf4	4	55527143	55532466	742.1172684	-0.306322312	0.000825173	0.046201106
ENSMUSG000000038886	Man2a2	7	80349097	80371375	2749.176919	-0.249114404	0.000824549	0.046201106
ENSMUSG000000038366	Lasp1	11	97799000	97838764	6961.310983	-0.193621576	0.000833814	0.046523372
ENSMUSG000000042520	Ubap2l	3	90000140	90052628	3661.790285	-0.220162162	0.00084301	0.046874255
ENSMUSG000000067786	Nnat	2	157560078	157562522	132.6289112	-0.359779712	0.000848334	0.047008181
ENSMUSG000000098557	Kctd12	14	102976581	102982637	1888.203368	-0.255471203	0.000856458	0.047295859
ENSMUSG000000037071	Scd1	19	44394451	44407709	5248.92282	-0.261679563	0.00087855	0.048350238
ENSMUSG000000028583	Pdpm	4	143267431	143299564	2865.303963	-0.266596927	0.000882907	0.048424761
ENSMUSG000000015957	Wnt11	7	98835112	98855195	113.8365579	0.356846739	0.00090119	0.04918927
ENSMUSG000000047443	Fam132b	1	91366430	91374217	193.8134177	0.346524775	0.000902947	0.04918927

List of significantly mutated genes

Table 2 List of significantly mutated genes in liver tumours. Genes are ordered by the significance of observed mutations for each group: DEN-initiated tumours from *Ctcf* hemizygous and wild-type mice, and spontaneously occurring murine tumours. Each group also includes the top seven genes mutated in other rodent tumours: *Hras*, *Braf*, *Egfr*, *Kras*, *Apc*, *Ctnnb1*, and *Keap1* (unpublished data). For each group, the number of tumours in which each gene is mutated is stated and the cumulative mutation rate per kb of the gene is given. P-values are corrected for multiple testing across all genes using the Bonferroni method.

Gene name	Ensemble gene ID	Gene length (bp)	No. tumours mutated	Mutations / kb	Corrected p value
DEN-induced tumours - wild-type (n = 32)					
Braf	ENSMUSG00000002413	122227	20	7.657	0
Hras	ENSMUSG00000025499	4901	5	6.954	0
Ext2	ENSMUSG000000027198	161541	7	2.922	0
Egfr	ENSMUSG000000020122	165956	8	2.086	0
Ahnak2	ENSMUSG000000072812	30464	7	0.525	0
Adamts18	ENSMUSG000000053399	151613	12	3.197	0.045
Ctnnb1	ENSMUSG00000006932	31292	4	1.659	1
Kras	ENSMUSG000000030265	33541	3	4.280	1
Apc	ENSMUSG00000005871	101266	2	0.223	1
Keap1	ENSMUSG00000003308	9632	0	0	1
DEN-induced tumours - <i>Ctcf</i> hemizygous (n = 34)					
Braf	ENSMUSG00000002413	122227	21	8.040	0
Hras	ENSMUSG00000025499	4901	6	8.345	0
Kras	ENSMUSG000000030265	33541	4	5.706	0
Hap1	ENSMUSG00000006930	8802	5	2.508	0
Ros1	ENSMUSG00000019893	149524	18	2.497	0.023
Ctnnb1	ENSMUSG00000006932	31292	4	1.659	1
Apc	ENSMUSG00000005871	101266	4	0.447	1
Egfr	ENSMUSG000000020122	165956	3	0.782	1
Keap1	ENSMUSG00000003308	9632	2	1.057	1
Spontaneous tumours (n = 6)					
Ctnnb1	ENSMUSG00000006932	31292	4	2.074	0
Braf	ENSMUSG00000002413	122227	0	0	1
Hras	ENSMUSG00000025499	4901	0	0	1
Egfr	ENSMUSG000000020122	165956	0	0	1
Kras	ENSMUSG000000030265	33541	0	0	1
Apc	ENSMUSG00000005871	101266	0	0	1
Keap1	ENSMUSG00000003308	9632	0	0	1